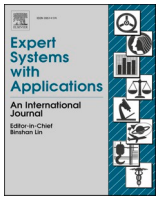
Expert Systems With Applications 237 (2024) 121490

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Modeling climate change impacts on vector-borne disease using machine learning models: Case study of *Visceral leishmaniasis* (Kala-azar) from Indian state of Bihar

Shubham Kumar [1], Aman Srivastava [1,2], Rajib Maity [*,3]

Department of Civil Engineering, Indian Institute of Technology (IIT) Kharagpur, Kharagpur 721302, West Bengal, India

ARTICLE INFO

Keywords:
Endemic
Sandfly
Neglected Tropical Diseases
World Health Organization
Data mining
Disease mapping
Artificial Neural Networks
Vector control strategy
Healthcare

ABSTRACT

Visceral leishmaniasis or Kala-azar (KA) is a Vector-Borne Disease (VBD) that remains the second-largest parasitic killer across the globe (mortality rate: 75–95%). More than 60% of KA cases originate in South Asia, wherein India accounts for 2/3rd of the cases, and Bihar, a state in India, alone accounts for more than 50% of the Indian cases. Past studies suspected climate change vulnerabilities as a driving cause of KA outbreaks. The VBDs-based epidemic prediction systems have been developed to mitigate recurrent outbreaks; however, Machine Learning (ML) based approaches still need to be explored for modeling changing climate impacts on KA cases. This study, for the first time, develops a Radial Basis Function (RBF) kernel-based Support Vector Regression (SVR), hereinafter RBF-kernel-based-SVR model for the most-affected endemic districts of Bihar (northern-India), using the data from 2016 and 2021. Forward selection, backward elimination, and stepwise regression procedures were adopted while selecting influential climatic variables, followed by the *k*-fold cross-validation technique and, then, the RBF-kernel-based-SVR algorithm for classification. Results suggested that temperature, wind speed, rainfall, and population density significantly contributed to the KA outbreaks. This study also developed Multiple Linear Regression (MLR) and Multilayer Perceptron (MLP) models to compare SVR with other classification models. Findings indicated that the proposed RBF-kernel-based-SVR model [Correlation Coefficient (CC) = 0.82, Root-Mean-Square Error (RMSE) = 12.20, and Nash–Sutcliffe Efficiency (NSE) = 0.66] outperformed MLR (0.81, 14.20, 0.48) and MLP (0.81, 12.95, 0.61). Study recommends using the RBF-kernel-based-SVR model as a quick and efficient model capable of detecting KA cases with high predictability even under limited data availability. Such models can assist public health authorities, given monitoring KA spread, learning the climate impacts of outbreaks, and ensuring timelier health services.

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) reported anthropogenic Greenhouse Gas (GHG) emissions as a primary cause of rapid global warming (IPCC, 2021; Srivastava et al., 2022). One of the profound long-term implications of Earth warming has been observed as a challenge while controlling Vector-Borne Diseases[4] (VBDs) (Rocklöv & Dubrow, 2020). Despite non-climatic drivers, past studies have frequently reported the negative consequences of climate change drivers on the transmission dynamics, geographic spread, and re-emergence of VBDs vide numerous pathways, such as through humans themselves, non-human hosts, pathogens, and vectors (Caminade et al., 2019; de Angeli Dutra et al., 2023; Franklinos et al., 2019; Rocklöv & Dubrow, 2020; Wilson et al., 2020). For example, ectotherms, vectors belonging to the cold-blooded category responsible for causing VBDs, are known to perform better in a warming climate (Fouque & Reeder, 2019). A few

* Corresponding author.
E-mail addresses: amansrivastava1397@kgpian.iitkgp.ac.in (A. Srivastava), rajib@civil.iitkgp.ac.in (R. Maity).
[1] Both authors have contributed equally to this work and shared the first authorship.
[2] ORCID: https://orcid.org/0000-0001-9253-3485.
[3] ORCID: https://orcid.org/0000-0001-5631-9553.
[4] Vector-Borne Diseases (VBDs) refers to infectious diseases transmitted to humans or animals through the bites of infected vectors, such as mosquitoes, ticks, fleas, sandflies, or other arthropods. These diseases include malaria, dengue fever, Zika virus, Lyme disease, Chagas disease, and many others. VBDs are a significant global public health concern, particularly in regions where vectors thrive, and environmental conditions favor their transmission.

https://doi.org/10.1016/j.eswa.2023.121490
Received 27 June 2023; Received in revised form 31 August 2023; Accepted 5 September 2023
Available online 14 September 2023

instances of warming climate include the increase in the number of warm days and nights, increased events of heatwaves, rising sea levels, greater warming over land than over the oceans, greater warming in winter than in summer, greater warming in nighttime than in the daytime, etc (IPCC, 2021; Srivastava et al., 2022). They have altogether accelerated evaporation rates, thereby intensifying the overall hydrological cycle (Abbott et al., 2019). The direct impacts of the disturbed hydrological cycle have been reflected as a rise in extreme precipitation events, which have more or less yielded a suitable environment/climate for the vectors to subsist and develop (Okoro et al., 2023; Rocklöv & Dubrow, 2020). The epidemiological triangle of VBDs, which comprises the host, pathogen, and transmitting agent, is influenced by climate factors. Temperature, precipitation, and humidity at various stages of their development exert distinct effects (Okoro et al., 2023; Wilcox et al., 2019). Among the various categories of VBDs, the World Health Organization (WHO) has recognized Neglected Tropical Diseases[5] (NTD) as a significant global health concern. These NTDs encompass a range of infectious diseases, including Dengue, Chikungunya, *Lymphatic filariasis*, Schistosomiasis, Onchocerciasis, Chagas disease, *African trypanosomiasis*, and Leishmaniasis (https://www.who.int/neglected_diseases/diseases/summary/en/, accessed April 2022). There are broadly three forms of Leishmaniasis, viz., *Cutaneous leishmaniasis*, *Mucocutaneous leishmaniasis*, and *Visceral leishmaniasis* (VL) or also called Kala-azar[6] (KA) in India (Bhunia & Shit, 2020a; Yadav et al., 2023). The present study concerns VL or KA (henceforth KA) disease.

KA is a slowly progressing indigenous disease caused by the protozoan parasite – *Leishmania donovani*. The primary vector of KA is *Phlebotomus argentipes* (sandfly), and the primary non-human reservoir (competent) hosts are rodents, dogs, and other mammals (Bhunia & Shit, 2020a; Rocklöv & Dubrow, 2020). Due to its high mortality rate of 75–95%, KA has been reported as the second-largest parasitic killer across the globe after Malaria (Bhunia & Shit, 2020a; Gil et al., 2020). Currently, there are 88 endemic countries across the globe affected by KA disease, of which 72 are underdeveloped or developing (Ahmed et al., 2020; Karunaweera & Ferreira, 2018). More specifically, around 200,000 to 400,000 new KA cases and 20,000 to 40,000 deaths per annum are consistently reported globally (Ahmed et al., 2020). Of these, 90% of the cases are observed concentrated across countries such as India, Bangladesh, Brazil, Ethiopia, Nepal, and Sudan. In addition, more than 60% of the cases originate exclusively in South Asia (Bhunia & Shit, 2020a).

India is a rapidly developing nation in the Asian continent; located in Southern Asia. The Indian subcontinent accounts for over two-thirds of the world's KA cases. About 54 districts from different states of India (for example, Bihar, Jharkhand, Uttar Pradesh, and West Bengal) are reported endemic due to KA cases, which have consequently risked over 130 million population. Bihar is one of the endemic states located in the eastern part of a subtropical region of India. In addition, Bihar is a climate-sensitive state due to its geographical setting, hydro-meteorological uncertainties, dense rural population, and high poverty level (Kumar et al., 2020; Kumar et al., 2022; Mahajan et al., 2023; Tesfaye et al., 2017). More than 50% of the KA cases in India are reported from Bihar. As far as endemic districts are concerned, 33 out of 54 (>61%) endemic districts are located in Bihar (Kumar et al., 2020). Considering the seriousness of the KA disease across the globe and in India (Bihar state in specific), the present study ascertained to extend the scientific investigation of KA incidences in parts of Bihar.

In the context of changing climate, there is a need to quantify the degree of spread of KA disease and other VBDs. Such practice may allow for a better account of the severity of the diseases via clear identification of the endemic regions and by devising the required support system or model (vector elimination schema). Investigations in these areas can additionally aid in focused policy-making, revising grassroots governance for disease control, and developing decentralized disease eradication models. In the recent decade, applications of Machine Learning (ML)-based model development for epidemiological data analysis has become widespread (Alfred & Obit, 2021; Joshi & Miller, 2021). ML algorithms are non-parametric and provide an empirical approach for conducting regression for non-linear systems (such as KA cases) involving a few variables to thousands of variables. Hence, ML applications find their strong applicability in cases with limited theoretical knowledge but a large number of observations. Some of the widely used ML algorithms include Support Vector Machines (SVMs), Artificial Neural Networks (ANNs) [such as Multilayer Perceptron (MLP)], Decision Trees (DTs) [such as Random Forest (RF), CART, XGBoost, and Regression Trees (RT)], k-Nearest Neighbors (kNN), Genetic Algorithms (GA), Fuzzy Logic (FL), and Topic Modeling (TM) (Ngiam & Khor, 2019; Sarker, 2021; Elbeltagi et al., 2022, 2023a, 2023b; Pande et al., 2022). Joshi and Miller (2021), in their extensive review of 120 papers on mosquito control analysis, brought to light the advancements made in applications with supervised and unsupervised ML approaches with ANNs, DTs, SVMs, GA, FL, and TM. Alongside limitations of the ML algorithms, the authors in this study discussed their unexplored potential in mitigating challenges of VBDs, such as through the development of an open-source ML pipeline, citizen science, and crowd-sourced data analysis.

Many recent studies are navigating to ML-based model development as one of the alternatives to generalize important trends of changing climate on the VBDs. Scavuzzo et al. (2018) developed and evaluated SVM, ANN, kNN, and DT regressors for temporal modeling of the oviposition activity of a specific vector responsible for causing Chikungunya, Dengue, and Zika viruses in Latin America. Their study found better modeling performance for ML algorithms against linear models while capturing anomalies. In another study by Tapak et al. (2019), the authors developed an Autoregressive Integrated Moving Average (ARIMA) model, SVM, RF, and ANN models for outbreak detection of influenza-like illness. Their study found RF as the best model and ANN as a better model for outbreak detection than linear models like ARIMA. Furthermore, Jimenez et al. (2020) developed Gaussian Process Regression (GPR), SVM, kNN, RF, Linear Regression (LR), and MLP models to predict antibiotic resistance outbreaks. Their study evaluated SVM as the best model for multivariate time series prediction. All such (aforesaid) investigations provide strong foundations for public health care and management using ML models, especially when controlling disease outbreaks. Again to accomplish this, Alfred and Obit (2021) conducted review research for all the papers employing ML models between 2010 and 2020 to discuss the ML algorithms, dataset, and performance measurements in predicting and detecting deadly infectious diseases. Key findings from their investigation revealed that (1) meteorological and epidemiology data are the most useful datasets for studying disease trends, and (2) SVM and ANN models provided better performance as compared to other linear and non-linear ML approaches

[5] Neglected Tropical Diseases (NTDs) are a diverse group of diseases that primarily affect populations in tropical and subtropical regions of the world, especially in low-income communities with limited access to healthcare. These diseases are often chronic and disabling, causing significant morbidity and mortality. NTDs include conditions such as lymphatic filariasis, schistosomiasis, soil-transmitted helminthiasis, onchocerciasis (river blindness), and others. Despite their substantial burden on affected populations, NTDs have historically received limited attention and resources for research, prevention, and treatment.

[6] Kala-azar is a severe form of leishmaniasis that primarily affects internal organs, including the spleen, liver, and bone marrow. Once the parasite enters the body, it multiplies within the immune cells, leading to a systemic infection. The initial symptoms of Kala-azar may include intermittent fever, fatigue, and weight loss. Individuals may experience persistent or prolonged fever as the disease progresses, often accompanied by chills and night sweats. Other common symptoms include enlargement of the spleen and liver (hepatosplenomegaly), which can cause abdominal discomfort and a visibly swollen abdomen. Anemia, resulting from the destruction of red blood cells, is also a characteristic feature of Kala-azar.

(the present study has followed both the recommendations).

The ML algorithms have demonstrated their wide-scale applicability while dealing with VBDs in general (as apparent from the previous discussions); however, it is imperative to highlight here that there are no research available for developing ML models for capturing insights for KA cases in specific. Since the applications of ML algorithms in medical sciences, more specifically in epidemiology, is an emerging field, the present study takes the opportunity to extend the ongoing investigations for unexplored KA disease time-series analysis using ML models. In light of this research gap, the present study aimed:

- To establish correlations between KA cases, population density, and climate variables, thereby ascertaining the influence of climate change on KA incidences;
- To develop Radial Basis Function (RBF) kernel-based Support Vector Machine (SVM) model for KA disease prediction in parts of Bihar state (India) and compare the results with Multiple Linear Regression (MLR) and Multilayer Perceptron (MLP) models;
- To compare the accuracy and stability of these models; and
- To determine which were the best outcomes based on the prediction accuracy vide the best combination of the input variables.

The key contribution of this paper is the development of the ML models, particularly in response to predicting the Kala-azar outbreaks and considering non-seasonal changes in climatic factors in one of the worst-affected study areas across the globe (i.e., endemic parts of Bihar state in India). Hence, the present findings on KA cases concerning climate change can be of immediate usage to the Health Ministry of Bihar, broadly in terms of revising grassroots governance for disease control and eradication. In addition, the presented findings and developed methodological framework, in general, can also be pertinent in a similar endemic site across the KA-affected countries.

2. Materials and methodology

2.1. Study area description

The Bihar state is located in the eastern region of India, specifically in the subtropical region of the temperate zone. The Ganges River, which flows from west to east, divides the Bihar plain into two unequal halves i.e., north Bihar and south Bihar. The present study sites – Muzaffarpur and Saran districts are located in west part of northern Bihar and share the great Indo-Gangetic plains (Fig. 1). The selection of these two districts out of the 33 endemic districts of Bihar is ascertained based on the large number of cases being reported from the west part of the northern Bihar (from where the current study areas have been shortlisted; refer to Fig. 4) majorly comprising Saran, Muzaffarpur, Siwan, Gopalganj, East (E) Champaran, Sitamarhi, and Vaishali.

Muzaffarpur district is located between 25°54′N and 26°23′N latitude and 84°53′E and 85°45′E longitude (see Fig. 1). It is situated at an altitude of 170 feet above mean sea level and occupies an area of 3,172 square kilometers (km^2). The study sites fall in the subtropical temperate zone with a humid subtropical climatic type (more description in Sect. 3.1). There are two administrative subdivisions (east and west) of the district, comprising a total of 1,811 villages under 385 Gram Panchayat (village-level governance body; third-tier of Government, after State and

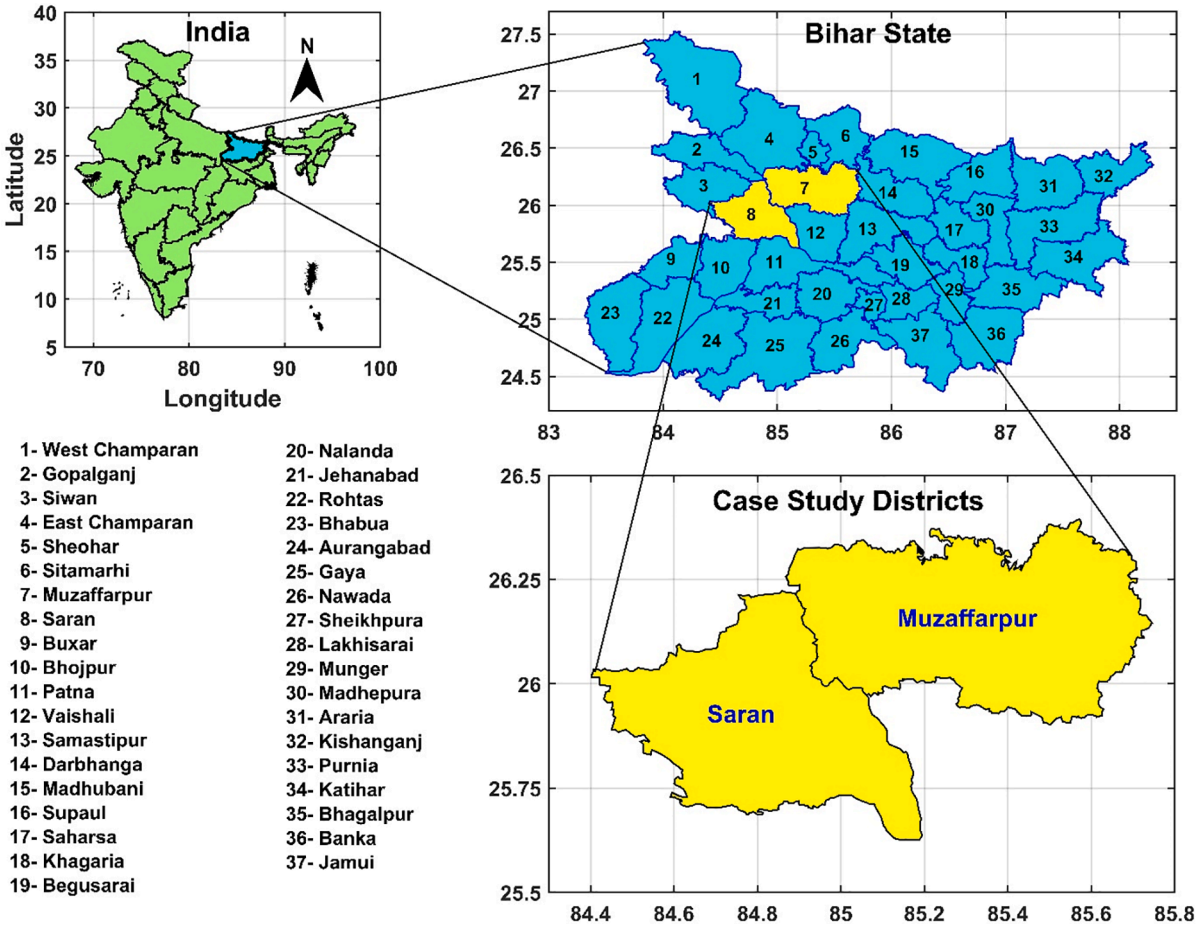


Fig. 1. Geographical map of the two selected endemic regions viz., Muzaffarpur and Saran districts in the Bihar state of India.

Central-level, in India) (CGWB-Muzaffarpur, 2013). Muzaffarpur has an urban population of 469,896 people (~10% of the district population) and a rural population of 4,308,714 people (~90% of the district population) and constitutes 4.61% of the total population in Bihar state with a population growth rate of 28.14% between 2001 and 2011. (Census of Muzaffarpur, 2011).

The western boundary of Muzaffarpur is shared with the second study site – the Saran district. It is located between 25°36′N and 26°13′N latitude and 84°24′E and 85°15′E longitude (see Fig. 1). It is situated at an altitude of 118 feet above mean sea level (since the district is located closer to the Ganga basin, the altitude is lesser than Mazuffarpur) and occupies an area of 2,641 km^2. Like Muzaffarpur, Saran also shares a humid subtropical climate (more description in Sect. 3.1). There are three administrative subdivisions (Chapra, Marhaura, and Sonpur) of the district, comprising a total of 1,807 villages under 330 Gram Panchayat (CGWB-Saran, 2013). Saran has an urban population of 352,045 people (~9% of the district population) and a rural population of 3,591,053 people (~91% of the district population) and constitutes 3.8% of the total population in Bihar state, with a population growth rate of 21.64% between 2001 and 2011 (Census of Saran, 2011).

2.2. Clinical data sources and acquisition

The confirmed KA cases for a period of seven years ranging from January 2016 to July 2021 for the Bihar state were obtained from the National Vector Borne Disease Control Programme (NVBDCP), Patna, Bihar. The dataset contains month-wise KA cases at both district and block scales. The confirmed cases of Kala-azar (KA) represent recorded incidences of *Visceral leishmaniasis* that have been verified and documented by various hospitals and Primary Health Care (PHC) centers. The data is then transferred to the NVBDCP, a specialized directorate responsible for framing technical guidelines, policies, and budgeting and planning the logistics to guide the states in implementing program strategies. Fig. 2 shows the month-wise variations in KA cases in Muzaffarpur and Saran districts from January 2016 to July 2021. It can be observed that the number of cases in the recent period (post-2018) across both districts has shown a decreasing trend. It is hypothesized that the vagaries of changing climate, as well as socio-economic conditions, must be strongly correlated with the KA outbreaks. Therefore, it is imperative to investigate the reasoning behind the recent fall and, in general, the high number of cases, as compared to other parts of India and the world, under both changing climate and socio-economic developments. Considering this, the present study considered KA cases as the dependent variable where the corresponding time-series data of monthly KA cases are log-transformed to stabilize the variance (detailed in Sect. 2.4).

2.3. Climatic data sources and acquisition

The climate datasets were collected from the repository of the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis product (ERA5, https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, accessed September 2023). The dataset contains observational data for a period of six years (2016–2021). The data records include independent variables, such as total precipitation (Rain), average temperature (T_avg), maximum temperature (T_max), minimum temperature (T_min), wind speed (WS), and relative humidity (RH). The details of each dataset, as obtained from ERA5, are provided in Table 1. This study has considered only the month-wise record of climate variables (Rain, T_avg, T_max, T_min, WS, and RH), population density (P_density), and KA cases (KA_cases). This is achieved by spatially averaging each dataset across the study sites and by converting it to a monthly scale before further processing.

2.4. Model development

A flow diagram summarizing the complete model development (for SVR, MLR, and MLP) is shown in Fig. 3. Different components and steps used for developing the models are explained in the following subsections.

2.4.1. Correlation analysis

The Pearson Correlation Coefficient (PCC) is a statistical measure that quantifies the strength and direction of the linear relationship between two variables (also refer to Sect. 2.5). It ranges from −1 to +1, where a value of −1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation. However, in the present study, there were multiple variables (Rain, T_max, T_min_Tavg, RH, P_density, and KA_cases) such that the objective was to establish a correlation analysis between all these variables in a matrix data structure called a correlation matrix (also called correlation heatmap) (Babicki et al., 2016; Hardin et al., 2013). The heat map visually represents the correlation coefficients between the dependent variable (KA cases) and each independent variable (Rain, T_avg, T_max, T_min, WS, RH, and P_density). Each cell in the heatmap corresponds to the correlation coefficient between two variables, with color gradients

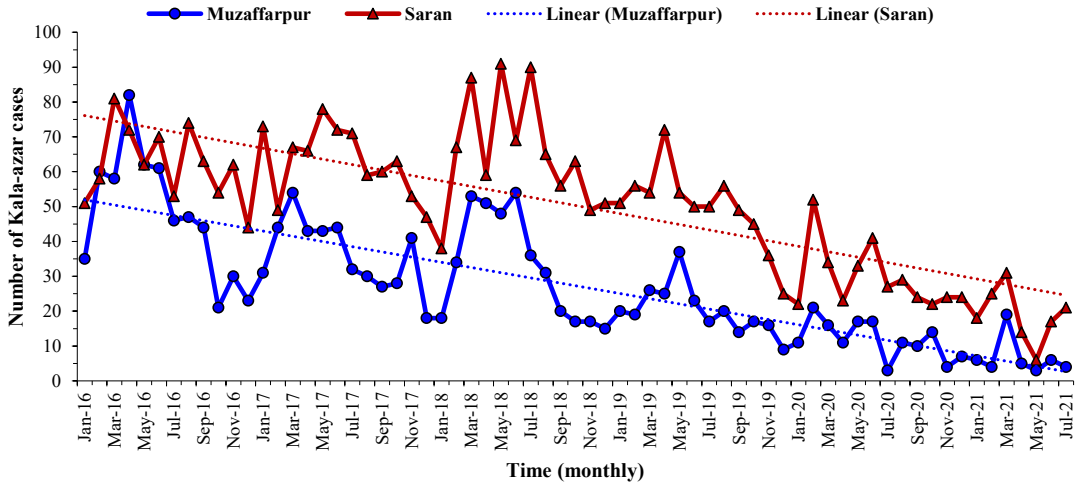


Fig. 2. Month-wise variation in *Visceral leishmaniasis* (VL) also called Kala-azar (KA) cases in Muzaffarpur and Saran districts in the Bihar state of India from January 2016 to July 2021 [Source: National Vector Borne Diseases Control Program (NVBDCP), Government of India].

Table 1
Details of the climatological dataset obtained from the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis product (ERA5).

Dataset	Variables	Nomenclature	Spatial resolution	Vertical/pressure level	Units (as in ERA5)	Converted Units
ERA5-Land monthly averaged data	Total precipitation	Rain	0.25°×0.25°	Surface	mm/day	mm
	2-m temperature	T_avg	0.25°×0.25°	2-m above surface	K	°C
	2-m Maximum temperature	T_max	0.25°×0.25°	2-m above surface	K	°C
	2-m Minimum temperature	T_min	0.25°×0.25°	2-m above surface	K	°C
	10-m u-wind	WS	0.25°×0.25°	10-m above surface	m/s	–
	10-m v-wind	WS	0.25°×0.25°	10-m above surface	m/s	–
ERA5 monthly averaged data on pressure levels	Relative humidity	RH	0.25°×0.25°	1000 hPa	Percent (%)	–

indicating the strength and direction of the correlation. In the present study, these matrix tablets were established for the Muzaffarpur and Saran districts to ascertain which pairs of variables were most closely correlated and thereby identify relationships between variables. For this, colored cells were used instead of monochromatic representation, such that the cell color corresponds to the number of measurements that match the dimensional value.

2.4.2. Predictor selection using forward selection and backward elimination

The forward selection method introduces one variable to the model at a time. The initial variable in the variable addition procedure refers to the one with the highest degree of correlation with the independent variable. The second variable is the one that is included in the model after the first variable's effects have been adjusted. In addition, this variable has the second-greatest degree of correlation with the independent variable. The variable addition process is repeated until all variables have been added or until the final variable with an insignificant regression coefficient has been added, at this point, no significant variables are left to add. Backward elimination, unlike the forward selection method, starts with the whole model and eliminates one variable at a time. The first variable to be removed is the one with the least contribution to reducing the predictive error sum of squares. Assuming there are additional insignificant variables, the method begins by deleting the next most insignificant variable. The elimination process ends when all of the variables are significant or all but one have been eliminated (Lindsey & Sheather, 2010; Speiser et al., 2019; Xu & Zhang, 2001).

This study followed the aforementioned forward selection and backward elimination methods while selecting the most significant variables and eliminating the most insignificant variables, respectively. OLSRR (ordinary least squares regression) library in R studio was used for executing forward selection and backward elimination methods. Here, the variables were tested at lag 3, yielding five significant variables viz., KA_cases at lag 1, P_density at lag 0, P_density at lag 1, Rain at lag 0, and Rain at lag 1. In the next step, this study log-transformed the target variable [i.e., KA_cases was transformed to log(KA_cases)] at lag 3 for determining the influential variables. This step yielded seven significant variables viz., KA_cases at lag 1, T_avg at lag 0, T_avg at lag 1, T_min at lag 1, WS at lag 2, Rain at lag 0, and Rain at lag 1. Although P_density was not listed in this step, considering the high significance of P_density on KA_cases (as confirmed by the large value of population density in the study sites and was ascertained during the correlation analysis; refer to Sect. 3.2), P_desnity was also employed in the list of the significant variable.

2.4.3. Step-wise regression and the k-fold technique

In stepwise regression, a variable that was included in the model at the beginning of the process may be removed at the subsequent stages. The regression followed the same calculations as established for the forward selection and backward elimination methods. More specifically, the forward selection process was primarily followed in the stepwise regression process. However, the potential of eliminating a variable was evaluated at each stage, as in backward elimination. The number of variables preserved in the model was determined by the levels of significance anticipated for variable inclusion and exclusion (Lindsey & Sheather, 2010; Speiser et al., 2019; Xu & Zhang, 2001). This study finally adopted a significance level of 0.05 for including variables in the model and a significance level of 0.1 for excluding variables from the model.

In order to address the data insufficiency issue due to the limited study period of six years, this study employed the k-fold cross-validation technique (Pal & Patel, 2020; Refaeilzadeh et al., 2009) for model training and testing. The monthly data points from January 2016 to July 2021 were combined for both Muzaffarpur and Saran districts, resulting in a total of 134 data points (67 for Muzaffarpur and 67 for Saran). Following the methodology outlined in Section 2.4.2, the study utilized the log-transformed target variable [log(KA_cases)] and applied stepwise forward selection and backward elimination at a lag of 3 to determine the influential variables. This approach allowed analyzing the data, considering temporal dependencies, rigorously. As a result, the study identified eight significant variables: KA_cases at lag 1, T_avg at lag 0, T_avg at lag 1, T_min at lag 1, WS at lag 2, Rain at lag 0, Rain at lag 1, and P_density at lag 0. To perform the k-fold cross-validation, the data points were divided into three folds ($k = 3$). This division involved randomly splitting the data into training and testing sets, ensuring that each fold contained a proportionate dataset representation Specifically, 67% of the data were allocated for training and 33% for testing. During the training phase, the models were trained using the training dataset of two folds (approximately 89 data points). The models were then evaluated using the remaining fold (approximately 45 data points) during the testing phase. This process was repeated three times, each time using a different fold as the testing set, while the remaining two folds served as the training set. By employing the k-fold cross-validation technique, this study aimed to assess the performance and generalizability of the developed ML models while mitigating the potential bias introduced by using a single train-test split. This approach allowed for obtaining robust and reliable estimates of the model's predictive capabilities and assessing its effectiveness in predicting KA cases. Steps provided in Sects. 2.4.1 to 2.4.3 were followed for developing all three models – SVR, MLR, and MLP and are further detailed in Sects 2.4.4 to 2.4.7. It is imperative to highlight at this point that the model development was preceded by the standardization of both influential dependent and independent variables with a mean value of zero and a Standard Deviation (SD) value of one (except for the MLR model). This was done to enhance the

learning speed of the models during the training and testing period.

2.4.4. Support Vector Regression (SVR) model

As the primary aim was to develop an ML model for the Muzaffarpur and Saran districts, the present study developed an SVR model using the RBF kernel. The kernel function generally provides Support Vector Machines (SVMs) flexibility while implicitly mapping the data to a higher dimensional feature space. As a result, non-linear solutions in a lower-dimensional feature space get correspond to linear solutions in a higher-dimensional feature space. This allows for employing SVR for modeling the non-linear problems involving hydroclimatology (climate variables in the present case). As compared to many kernel functions, some of the earliest studies have identified RBF to be outperforming others (e.g., Dibike et al., 2001; Bray & Han, 2004; Liong & Sivapragasam, 2002; Choy & Chan, 2003; Maity et al., 2010). Moreover, recent studies have further affirmed the RBF credibility among many kernel functions (e.g., Ding et al., 2021; Gopi et al., 2023; Hekmatmanesh et al., 2020; Nguyen et al., 2021). Thus, the RBF-based SVR became the natural choice for the present study. Vapnik and others at AT&T Bell Laboratory were the first to lay a foundation for the present form of SVMs back in the early 1990s (Boser et al., 1992; Guyon et al., 1992). Since then, rapid transformation in SVMs happened that led to their first application in the late 1990s (Vapnik, 1998; 2000). For RBF kernel-based SVR model development, $\{x_i, y_i\}$ was considered the training set, ranging between $i = 1$ to n such that $x_i \in \chi^p$ indicates the p-dimensional input feature and $y_i \in \chi$ indicates system output. In this process, the first objective was to construct an activation function that logically defines the dependence of y on x [i.e., $y = f(x)$]. For this, a linear function was formulated for SVR, as shown in Equation (1), where w and b represent weight vector and bias, while $\psi(x)$ showing a non-linear mapping function. The function $\psi(x) : \chi^p \rightarrow \chi^h$ allows data to be mapped in a higher-dimensional space, converting the non-linear separable problem into a linearly separable problem. This can be explained as an optimization (*Min*) problem, as shown in Equation (2), where r shows random error and $\phi \in \chi^+$ shows a regularization parameter, which is used for optimizing the training errors and model's complexity in terms of their minimization.

$$y = w^T \psi(x) + b \tag{1}$$

$$Min\ J(w,\ r) = \frac{1}{2} w^T w + \frac{\Phi}{2} \sum_{i=1}^{n} r_i^2 \tag{2}$$

The next objective was to determine the optimal parameters that reduce the regression model's prediction error. This was achieved by selecting the optimal model, which showed reduced cost function and minimal e_i. The Lagrange function facilitated the selection, as shown in Equation (3). The solution of this equation was obtained by partially differentiating it with w, r, b, and μ, and the final equation of the SVR model was obtained, as shown in Equation (4). It is imperative to highlight here that $k(x, x_i)$ represents the kernel function such that Equation (5) shows the RBF kernel function, where σ represents the kernel function parameter of the RBF kernel.

$$L(w, r, b; \mu) = J(w, r) - \frac{1}{2} w^T w + \sum_{i=1}^{n} \mu_i \{w^T \psi(x_i) + b + e_i - y_i\} \tag{3}$$

$$y = f(x) = \sum_{i=1}^{n} \mu_i k(x, x_i) + b \tag{4}$$

$$k(x, x_i) = \exp\left(-\frac{1}{\sigma^2} ||x - x_i||^2 \right) \tag{5}$$

The regularization parameter ($\phi \in \chi^+$) allowed for minimizing fitting error and smoothening of the estimated function. To maximize the performance of the SVR model, the study optimized the magnitudes of ϕ and σ during the model calibration stage. In the model training stage, the grid-search technique using a cross-validation approach was employed for fine-tuning ϕ and σ. The output obtained from the resultant SVR model was then used for predicting the KA cases in the study sites.

2.4.5. Multiple linear regression (MLR) model

In order to compare the findings of the SVR model with other models, the MLR model was developed. Linear regression is the case of the relation between a dependent and a single independent variable. However, in the present case, the dependent variable (KA_cases) was observed to be dependent on several independent variables (Rain, T_max, T_min_Tavg, RH, and P_density). Dependence on multiple independent variables represented the case of multiple linear regression, and thus study focused on developing an MLR model (Andrews, 1974; Grégoire, 2014; Tranmer & Elliot, 2008). Equation (6) shows an expression for a general MLR model, where X_1, X_2, ..., X_p shows the independent variables, Y represents an independent variable, β_0, β_1, ..., β_p shows the unknown parameters, and ε shows the residual values.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{6}$$

Furthermore, while developing the MLR model, this study ensured that the assumptions of the linear regression were followed and data remained free from the multi-collinearity effect. As for the given n observations, a set of data may contain n observations of Y and corresponding n observations of p independent variables. Hence, Equation (6) can be rewritten in the algebraic form as Equation (7) and matrix form as Equation (8), where Y_i is the i^{th} observation of the dependent variable and $X_{i,\ j}$ is the i^{th} observation of the j^{th} independent variable. The least-square method was employed for determining the values of the parameters.

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} \tag{7}$$

$$[Y]_{n\times 1} = [X]_{n\times(p+1)} \cdot [\beta]_{(p+1)\times 1} \tag{8}$$

2.4.6. Multilayer Perceptron (MLP) model

Besides the development of the MLR model, this study also developed the MLP model for a comprehensive comparison of the results of SVR. MLPs are the feed-forward neural network models that provide a non-linear (static) mapping between an input layer (or vector) and a corresponding output layer. The structure of MLPs is composed of neurons called perceptions (or perceptrons); first proposed by Rosenblatt (1961). The n input features ($x = x_1, x_2, ..., x_n$) are received by a perceptron such that each input is associated with a weight ($w_1, w_2, ..., w_n$). Essentially, input features are required to be numeric to make them functional, else non-numeric input features are converted to numeric ones. The 3-layered MLP structure is the most widely used neural network model capable of approximating any continuous function. The 3-layers comprise (1) one input layer of perceptrons that functions to distribute input features to the first hidden layer, (2) one or more hidden layers such that the first hidden layer receives the input from the input layer and outputs them to further hidden layers, wherein each of which receives inputs from previous perceptron layers and provides as output to the next hidden layers, and (3) one output layer that receives the output of the last hidden layer as input feature (Car et al., 2020; Taud & Mas, 2018). Thus, the input and output layers are interconnected together through an intermediate hidden layer.

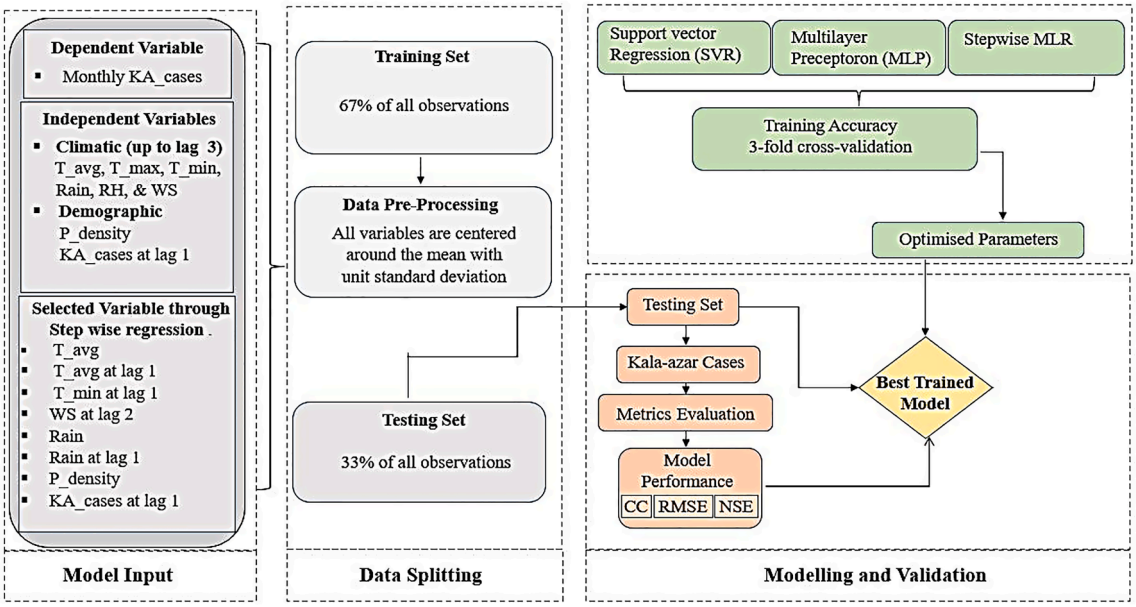


Fig. 3. Methodological flowchart comprising all three comparative models used in the study.

The MLP model developed for this study had five layers and was used for both the Muzaffarpur and Saran districts. Before the model development and as was discussed in Sect. 2.4.3, all the variables (dependent and independent) were standardized (mean = 0 and SD = 1) to facilitate fast learning of the model. After training and testing the model, the predicted results (in the log unit) were converted to the original scale. The working operation is described in Equations (9) to (13). The input layer [$U(x)$] of the MLP structure received the input feature using Equation (9). The output of the perceptron was computed using the activation function (f; usually a step function), as shown in Equation (10), such that the output was passed on to the next layer. Here, α is a threshold parameter. It can be understood here that perceptron allows for determining true or false conditions by examining $w_1 x_1 + w_2 x_2$, ... $+ w_n x_n - \alpha > 0$ such that $w_1 x_1 + w_2 x_2$, ... $+ w_n x_n - \alpha = 0$ represents the equation of hyperplane. The location of an output from a hyperplane indicated either 1 (if it is located above) or 0 (if located on or below) concerning the hyperplane. Adjustment in the weights of input features was then performed as a part of MLP or perceptron training such that hyperplane can be used to differentiate the training data. Following this, Equations (11) and (12) allowed for computation at hidden layer(s) [$H(x)$] and the output [$O(x)$], respectively, where S and G are activation functions (usually a sigmoidal function; differentiable for different layers), $b(1)$ and $b(2)$ are bias vectors, and $w(1)$ and $w(2)$ are weight matrix. In the current 5-layered MLP model, five hidden layers were used with one dropout layer at a dropout rate of 10% (to avoid overfitting). Furthermore, to generate a non-linear relationship, the "ReLu" activation function was used in hidden layers, as shown in Equation (13).

$$U(x) = \sum_{i=1}^{n} w_i x_i \tag{9}$$

$$f[U(x)] = \begin{cases} 1, if U(x) \rangle \alpha \\ 0, if otherwise \end{cases} \tag{10}$$

$$H(x) = S[b(1) + w(1)x] \tag{11}$$

$$O(x) = G[b(2) + w(2)H(x)] \tag{12}$$

$$ReLu = \begin{cases} x, if x > 0 \\ 0, if x \leq 0 \end{cases} \tag{13}$$

2.5. Model performance evaluation

The performance of the models developed for assessing KA cases outbreak in Muzaffarpur and Saran districts was conducted by using three standard statistical performance metrics, namely Correlation Coefficient (CC) (Pearson, 1896), Root-Mean-Square Error (RMSE) (Willmott & Matsuura, 2005), and Nash–Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970). The CC measures how well the model matches experimental data, as shown in Equation (14). The values of CC range from −1 to +1. A value of 1 indicates a perfect positive correlation, implying that the variables under consideration exhibit a strong linear relationship in the same direction. On the other hand, a value of −1 represents a perfect negative correlation, indicating a strong linear relationship, but in the opposite direction. Values closer to 0 suggest a weaker or no linear relationship between the variables. RMSE statistics represent the root mean square deviation of modeled values from the observed values of the time series, as shown in Equation (15). RMSE values range from 0 to positive infinity, with a lower value indicating better performance. In other words, the goal is to minimize the RMSE, with values closer to 0 considered more favorable. NSE is a normalized statistic that allows estimating the relative values of the residual variance while comparing with measured data variance, as shown in Equation (16), indicating the wellness of the plot (observed versus modeled data) concerning a 1:1 line. NSE value ranges between $-\infty$ and + 1 such that the output of 1 is considered the most accurate model (modeled data matching perfectly with observed data), the output of 0 is considered to have model predictions as accurate as the mean of the observed data, and the output between $-\infty$ and 0 is considered to have the mean of the observed data as a better predictor than the mean of the modeled data. In Equations (14), 15, and 16, O and P represent observed and modeled or simulated values for an i^{th} dataset; O_{Avg} and P_{Avg} represent the average or mean values of observed and simulated values, and N represents the number of observations. As described before, the simulation was performed through a k-fold (here, $k = 3$) cross-validation technique such that the mean values of CC, RMSE, and NSE across all three k-folds were obtained for model training and testing stages. Graphical representations of k-fold were developed separately for SVR, MLR, and MLP models for quantifying their performance using MATLAB.

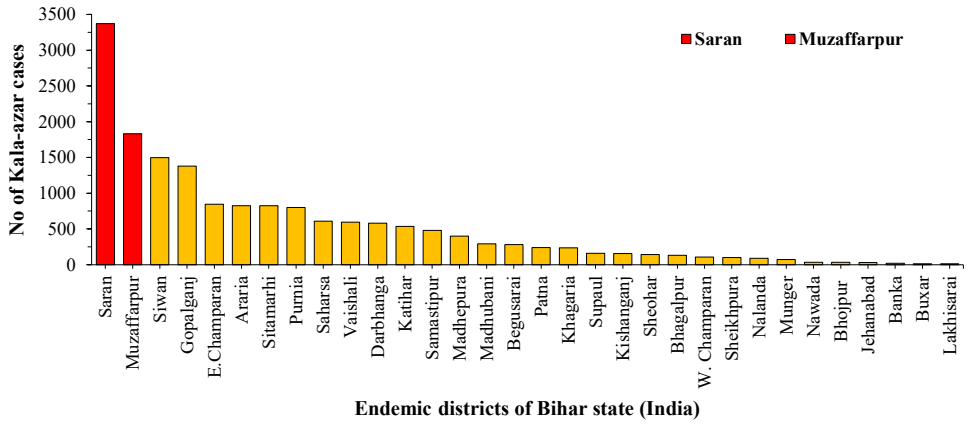


Fig. 4. Endemicity of Muzaffarpur and Saran districts to *Visceral leishmaniasis* (VL), also called Kala-azar (KA), shows the total number of cases being reported from January 2016 to July 2021 [Source: National Vector Borne Diseases Control Program (NVBDCP), Government of India].

$$CC = \frac{\sum_{i=1}^{N}(O_i - O_{Avg})(P_i - P_{Avg})}{\sqrt{\sum_{i=1}^{N}(O_i - O_{Avg})^2 (P_i - P_{Avg})^2}} \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2} \tag{15}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(O_i - O_{Avg})^2} \tag{16}$$

3. Results

3.1. Implications of climate change and population density on Kala-azar incidences

This study presents the findings on the frequency and pattern of KA incidences (in the current paragraph), trends and variations in climate variables (in the next paragraph), and population dynamics (in the last paragraph) across urban and rural settings. Taking into account the endemicity of KA cases reported from January 2016 to July 2021 (Fig. 4), it is evident that the districts located in the western part of northern Bihar, including Muzaffarpur and Saran, collectively accounted for 62% (10,342 cases) of the total reported cases in the state of Bihar. Remarkably, Muzaffarpur and Saran districts exhibited the highest number of reported cases, contributing 31% (5,202 cases) of the total KA cases in Bihar. Despite the decreasing trend observed in KA incidences (Fig. 2), the investigation identified Muzaffarpur and Saran as suitable choices for developing ML models to explore the correlation between KA and climate change dynamics. Moreover, Fig. 4 demonstrates their substantial contribution to the overall KA burden in Bihar and their suitability for representing the broader patterns and dynamics of the disease in the region. Imperatively, the decreasing trend observed in KA incidences does not diminish the significance of studying these districts. Instead, it provides an opportunity to investigate the underlying factors contributing to this trend and further explore the role of climate change in shaping the transmission dynamics of KA.

Regarding climate and weather dynamics, as discussed, the present study areas are located in the subtropical temperate zone, and their climatic type is humid subtropical. In general, the entire north Bihar primarily falls under mild and dry winter (December to February), hot summer (March to Mid-June), subtropical monsoon (Mid-June to September), and subtropical post-monsoon (October and November). The long-term temperature in winter varies around 0–10 °C, while December and January are recorded as the coldest months. Whereas the long-term temperature in the summer varies around 35–45 °C; May is recorded as the hottest month. The western part of northern Bihar receives a long-term annual average rainfall of around 1,040–1,450 mm (mm) (Kumar et al., 2020; Tesfaye et al., 2017). Considering the present study period (2016–2021), Fig. 5 shows the month-wise variations in rainfall, average temperature, and relative humidity in Muzaffarpur and Saran districts. The average monthly rainfall in Muzaffarpur is 115 mm, and in Saran is 107 mm, such that the peak is attained in August and/or September (Fig. 5a). The average monthly temperature in Muzaffarpur and Saran is recorded at 25.7 °C and 25.9 °C, respectively. The maximum average monthly temperature may increase to 33.9 °C and 34 °C, and the minimum average monthly temperature may fall to 17.6 °C and 17.9 °C, respectively (Fig. 5b). In the case of month-wise relative humidity, it is 68.7% in Muzaffarpur and 68% in Saran (Fig. 5c). While in the case of month-wise wind speed, it is 6.3 km/h in Muzaffarpur and 5.8 km/h in Saran.

Imperatively, it can be summarized from Fig. 5 that both study sites recorded an abrupt increase in the monthly rainfall and relative humidity magnitudes in a mere six-year period against the long-term average values (as apparent from the slope magnitude (m) = +0.026 for rainfall and +0.0025 for relative humidity). On the contrary, the temperature recorded a marginally decreasing trend (m = −0.0003). As it is known that VBDs, including KA, spread in a warming climate, one of the reasons for their recent decrease in the present study site may be attributed to the declining mean temperature magnitudes. Nevertheless, since all the aforementioned climatological variables (including wind speed) are combinedly known to govern KA cases and also influence the climate significantly, they all were thus considered for model development.

Geomorphologically, the present study sites share their location in alluvial plains, wherein Ganga, Gandak, and Ghagra are the key drainage basins. Moreover, each of these study sites elucidates a large number of villages (>1,800) and a comparatively greater section of the population (>90%) inside villages, as compared to the urban region (refer to Sect. 2.1), thereby demonstrating the developing stage of the districts. Additionally, as evident from Fig. 6, the population density is noticed to increase steadily in both study areas since 2016. Muzaffarpur witnessed an increase of 13.2% in population density between 2016 (from 1,198 persons/km^2) and 2021 (to 1,356 persons/km^2), whereas Saran witnessed a rise of 16.2% in population density between 2016 (from 1,250 persons/km^2) and 2021 (to 1,452 persons/km^2). Apparently, the large and rapidly increasing population density (greater in Saran than Muzaffarpur), alongside its congestion in rural localities, is indicative of a possible reason behind the KA outbreaks in these regions. Hence, apart from climate variables, the population density was also considered as one of the independent variables (P_density) while studying the climate change influence on KA cases as well as during the development of models. However, before discussing model results, it would be essential to bring a correlation analysis between climate variables, population density, and KA cases in Muzaffarpur and Saran districts.

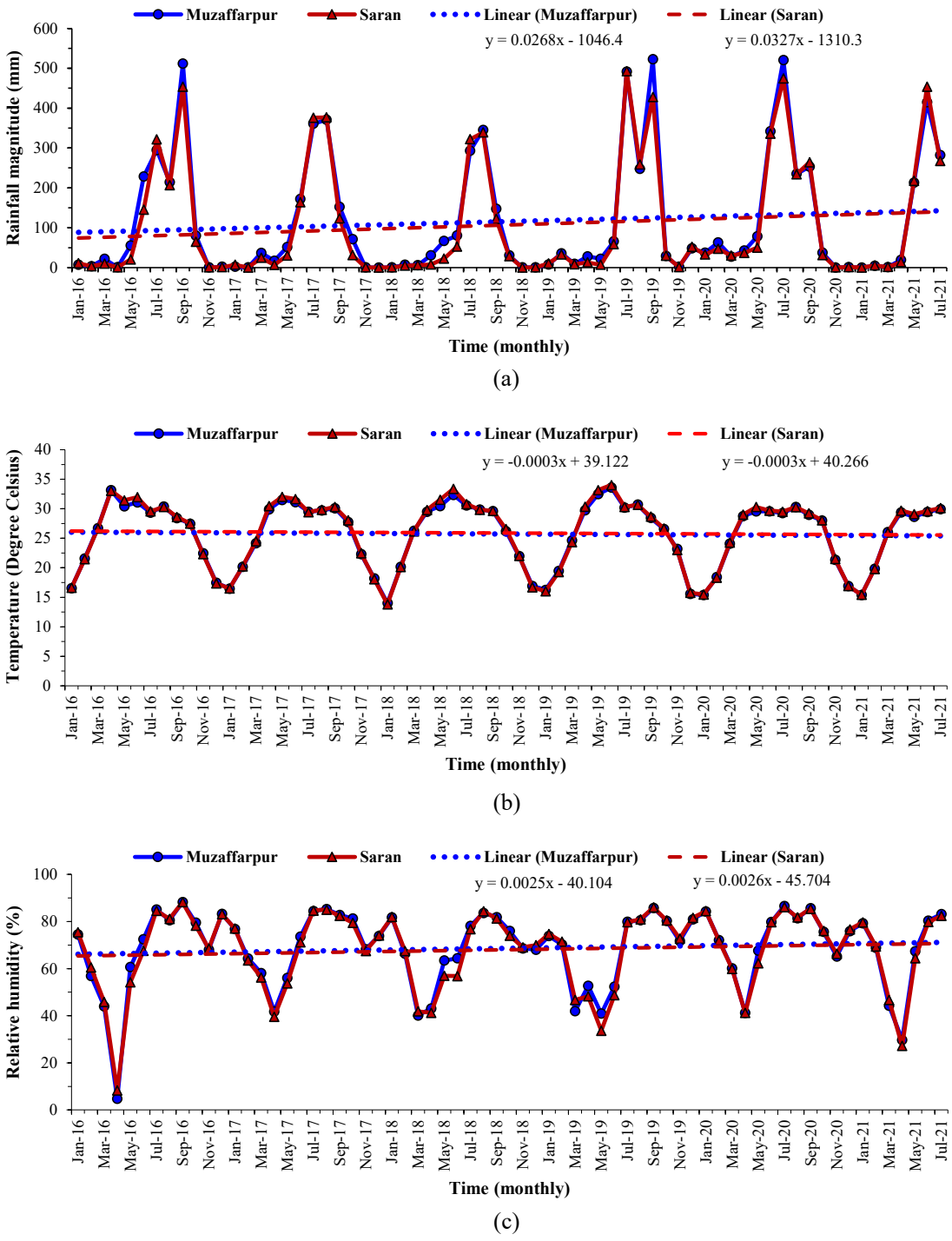


Fig. 5. Month-wise variation in rainfall, average temperature, and relative humidity in Muzaffarpur and Saran districts of Bihar state (India) between January 2016 and July 2021 [Source: Fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis product (ERA5)].

3.2. Correlation between climate variables, population density, and Kala-azar cases

The correlation matrices were developed individually for the Muzaffarpur and Saran districts. Figs. 7 and 8 show the correlation matrix depicting the correlation between KA cases, climate variables, and population density. In the case of Muzaffarpur, the analysis revealed positive correlations between the dependent variable KA_cases and four independent variables: T_max, T_min, T_avg, and WS. This implies that as these variables, such as temperature and wind speed, increase, KA_cases also tend to increase. Conversely, three independent variables, namely RH, Rain, and P_density, showed negative correlations with KA_cases, indicating that as these variables increase, the incidence of KA_cases tends to decrease. The analysis of PCC demonstrated that the correlation pattern between KA_cases and climate variables, as well as population density (P_density), remained similar for both Saran and Muzaffarpur districts. However, the strength of the correlation, as indicated by the PCC values, was observed to be slightly higher (~20%) in Saran compared to Muzaffarpur. In general, the correlation matrix indicated temperature and wind speed as highly correlated (or influential) with Kala-azar cases. It is imperative to re-highlight the findings obtained during the forward selection, backward elimination, and

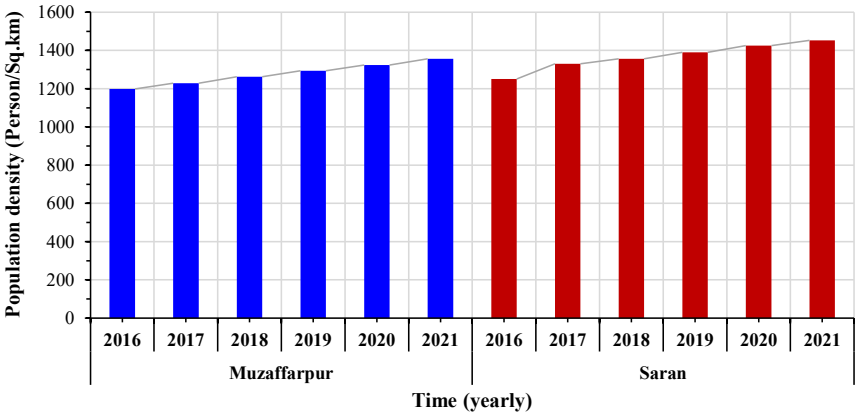


Fig. 6. Annual population density of Muzaffarpur and Saran districts of Bihar (India) for the study period January 2016 to July 2021 [Source: for Muzaffarpur, data was acquired from https://www.macrotrends.net/cities/21342/muzaffarpur/population, Accessed April 2022; for Saran, the population density was estimated using data acquired from Census of Muzaffarpur (2011), Census of Saran (2011), and aforementioned link because data is unavailable in the public domain and the primary source – 'Census of India – 2021' is yet to release its survey report; delayed due to global pandemic].

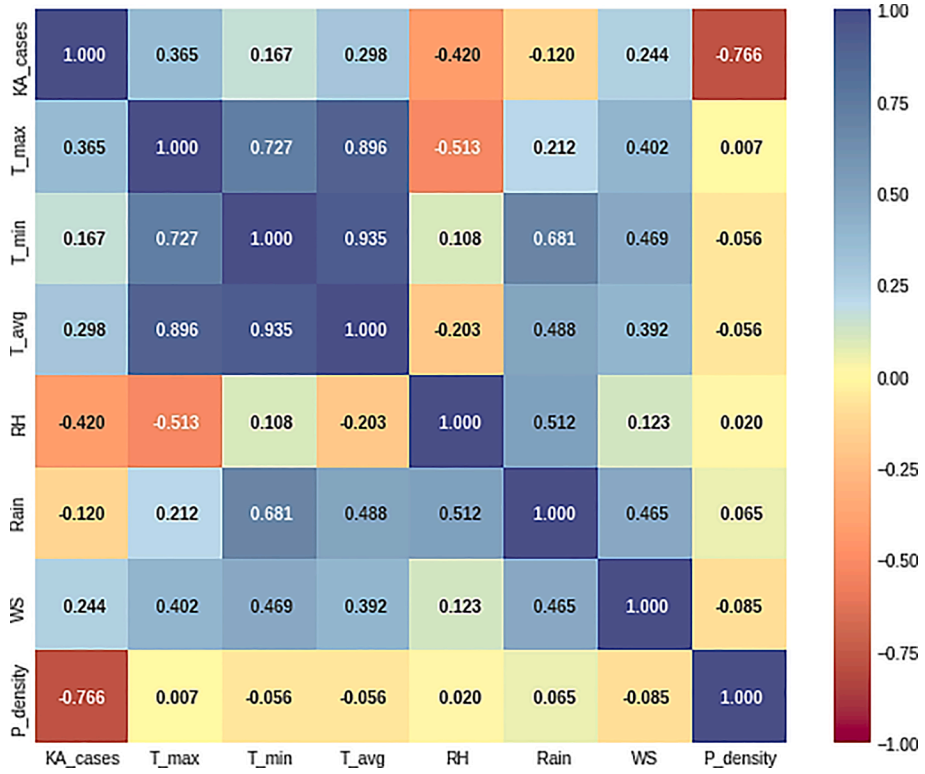


Fig. 7. Correlation matrix between dependent variable [KA_cases: Kala-azar cases] and independent variables [T_max: maximum temperature; T_min: minimum temperature; T_avg: average temperature; RH: relative humidity; WS: wind speed, and P-density: population density] for Muzaffarpur district of Bihar state (India).

stepwise regression methods. In the case of the first two methods, population density and rainfall were identified as significant variables that were replaced during log transformation with temperature, wind speed, and rainfall. The significant variables were identified as temperature, wind speed, rainfall, and population density during stepwise regression. To summarize, the findings from both correlation matrix and variable selection/elimination methods alongside regression, more or less, indicated that the variables – temperature, wind speed, rainfall, and population density were the influential variables that remained the cause behind aggravating KA cases in the study site.

Figs. 7 and 8 further show the correlation between the climate variables and population density. For both Muzaffarpur and Saran, a positive correlation was observed between KA_cases and maximum temperature (T_max), indicating that as temperature increases, KA_cases tend to increase as well, and vice versa. This positive correlation was also observed for T_min, T_avg, Rain, WS, and P_density, while RH showed a negative correlation. However, the strength of the correlation was found to be stronger in Muzaffarpur by around 25% on average. Additionally, T_min, T_avg, RH, and Rain showed negative correlations with P_density. Relative humidity and rainfall were positively correlated with WS, with Muzaffarpur showing a stronger correlation. Wind speed exhibited a negative correlation with P_density. In summary, temperature showed a stronger positive correlation with rainfall than relative humidity, and population density negatively correlated with wind speed and temperature.

Besides, it is essential to emphasize that the present study focused on a specific region in India, namely the Bihar state, with a particular focus on Muzaffarpur and Saran districts. These districts may have unique ecological and climatic conditions that differ from other regions within India and globally. Therefore, the findings demonstrated in the present and previous sections may not be directly generalizable to areas with distinct environmental characteristics. Furthermore, it is imperative to

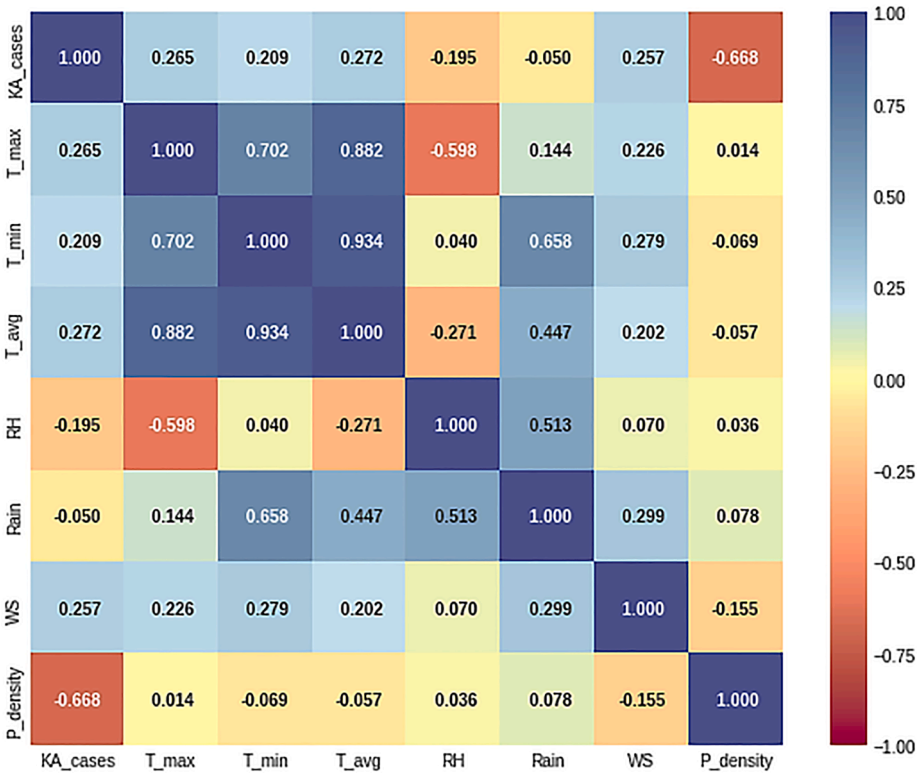


Fig. 8. Correlation matrix between dependent variable [KA_cases: Kala-azar cases] and independent variables [T_max: maximum temperature; T_min: minimum temperature; T_avg: average temperature; RH: relative humidity; WS: wind speed, and P-density: population density] for Saraan district of Bihar state (India).

recognize that VBDs, such as KA, are influenced by a complex interplay of various factors, including local ecological conditions, socio-economic factors, healthcare infrastructure, and human behavior. While this study strived to account for these factors, it is important to acknowledge that additional contextual variables may exist in different geographic regions that could influence the dynamics of KA. Moreover, in this study, correlation patterns were focused on rather than establishing statistical significance, in that the emphasis was on understanding the relationships and magnitudes of associations using forward selection, backward elimination, and stepwise regression methods, providing valuable insights into the variables of interest. Therefore, formal statistical tests for significance were not conducted. Notably, the omission of statistical significance testing is not uncommon, as evidenced by numerous recent works in the field (for example, Elbeltagi et al., 2023a, 2023b; Pradhan et al., 2022).

3.3. k-fold analysis of SVR, MLR, and MLP models in capturing Kala-azar cases

This section evaluates the potential of the proposed SVR-based model to capture KA cases in the Muzaffarpur and Saran districts and thereby compares it with MLR- and MLP-based models. Fig. 9 shows the comparisons between the observed and simulated values of KA cases through time series (left) and scatter plots (right) for the SVR model. A visual interpretation disclosed that the predictions were robust in terms of direction (above or below the normal) and the magnitude of KA cases. Fold-wise, during the training stage, fold-2 was observed to capture the peaks most accurately, followed by fold-1 and fold-3; however, on the contrary, during the testing stage, fold-3 became the most accurate capturer of peaks followed by fold-1 (remained comparable with the training stage) and fold-2. Overall, scatter plots indicated an acceptable correlation between the observed and simulated magnitude of KA cases for all folds. This demonstrates the reliable potential of the SVR-based model in the KA cases evaluation using the information on climate and population density.

Similarly, Figs. S1 and S2 (available in supplementary material) were prepared to compare observed and simulated KA cases' magnitudes through time series and scatter plots for MLR and MLP models, respectively. A visual interpretation of both models disclosed that the predictions were more or less promising, considering both the direction and the magnitude of KA cases. Concerning the MLR model, fold-wise inspection for the training stage revealed all the folds comparable while capturing the peaks; however, during the testing stage, fold-3 became the most accurate capturer of peaks, whereas both fold-1 and fold-2 remained comparable (coherent to the training stage). While in the case of the MLP model, fold-wise inspection for the training stage identified fold-2 as the best against the comparable performance of fold-1 and fold-3; the testing stage identified fold-3 as the best against the comparable performance of fold-1 and fold-2. Overall, scatter plots of both MLR and MLP indicated an acceptable correlation between the observed and simulated magnitude of KA cases for all folds. However, in the comparison, MLP was observed to perform better than MLR, and in general, SVR performed better than both MLR and MLP during the training and testing stages. Hence, it can be proclaimed with high significance that the SVR model successfully captured the complex relationship between different climatological variables and variations in KA cases for the present study site.

3.4. Comparison of SVR with MLR and MLP models for performance evaluation

The performances of the simulations for the SVR-based model were quantified through three performance metrics, namely CC, RMSE, and NSE. The magnitudes corresponding to each fold are summarized in Table 2. Additionally, the graphical representation for mean values of performance metrics is shown in Fig. 10. The mean performance of the SVR-based model during the training period ranged between 0.81 and 0.86 (CC), 11.01–12.83 (RMSE), and 0.65–0.73 (NSE), and that for the

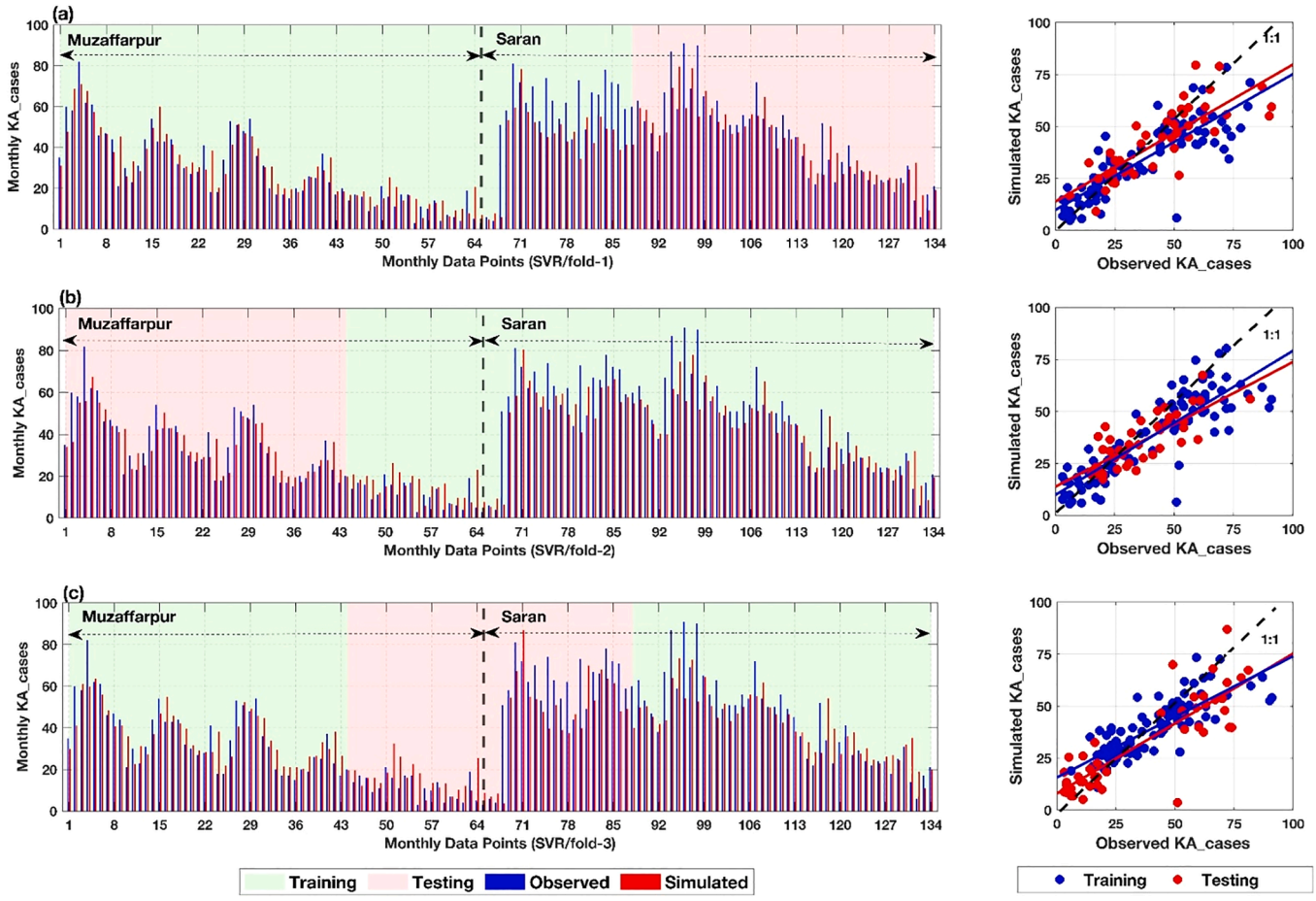


Fig. 9. Model performance in Kala-azar (KA) cases modeling for Muzaffarpur and Saran districts considering the Support Vector Regression (SVR) model. Observed and simulated monthly scale KA_cases are shown through time series (left) and scatter plots (right) for all 3 folds (a, b, and c). Training and testing periods are shown in the time-series plots for different folds. In the scatter plots, solid black lines show the 45° line (1:1; line of perfect simulation), and the other two lines show best-fit lines for the scatter plots.

Table 2
Fold-wise comparison of performance metrics [Correlation Coefficient (CC), Root-Mean-Square Error (RMSE), and Nash–Sutcliffe Efficiency (NSE)] between observed and simulated Kala-azar (KA) cases during training (Trn) and testing (Tst) periods obtained by Support Vector Regression (SVR), Multiple Linear Regression (MLR) and Multilayer Perceptron (MLP) models for the Muzaffarpur and Saran districts of Bihar (India).

Fold	Model	CC		RMSE		NSE	
		Trn	Tst	Trn	Tst	Trn	Tst
1	SVR	0.84	0.81	12.43	11.91	0.69	0.66
	MLR	0.84	0.78	12.48	19.67	0.69	0.07
	MLP	0.84	0.82	12.58	12.39	0.69	0.63
2	SVR	0.86	0.78	12.83	9.69	0.73	0.61
	MLR	0.84	0.78	13.79	9.91	0.68	0.59
	MLP	0.89	0.75	12.18	10.62	0.75	0.53
3	SVR	0.81	0.85	11.01	15.01	0.65	0.71
	MLR	0.83	0.89	10.59	13.02	0.68	0.78
	MLP	0.85	0.85	10.51	15.83	0.68	0.67
Mean	SVR	**0.84**	**0.82**	**12.09**	**12.20**	**0.69**	**0.66**
	MLR	**0.86**	**0.81**	**11.75**	**12.95**	**0.71**	**0.61**
	MLP	**0.83**	**0.81**	**12.29**	**14.20**	**0.68**	**0.48**

testing period ranged between 0.78 and 0.85 (CC), 9.69–15.01 (RMSE), and 0.61–0.71 (NSE). It can be inferred from the aforementioned statistical measures that the SVR model performance remained comparable during both the training and testing stages. Furthermore, it can be demonstrated that the model training was appropriate, resulting in no overtraining or undertraining. Hence, the developed SVR-based model could better simulate the KA cases in the present study region.

Similarly, the values corresponding to each fold for MLR and MLP models are summarized and graphically represented in the same tables and figures of SVR viz., Table 2 and Fig. 10, respectively. As mentioned before, the SVR-based model performance was also compared against MLR and MLP, keeping all other modeling conditions the same. The mean performance of the MLR-based model during the training period ranged between 0.83 and 0.84 (CC), 10.59–13.79 (RMSE), and 0.68–0.69 (NSE), and that for the testing period ranged between 0.78 and 0.89 (CC), 9.91–19.67 (RMSE), and 0.07–0.78 (NSE). While the mean performance of the MLP-based model during the training period ranged between 0.84 and 0.89 (CC), 10.51–12.58 (RMSE), and 0.68–0.75 (NSE), and that for the testing period ranged between 0.75 and 0.85 (CC), 10.62–15.83 (RMSE), and 0.53–0.67 (NSE). It can be inferred from the aforementioned statistical measures that the MLR and MLP model performance remained comparable during the training and testing stages. Furthermore, just like SVR-based model performance, it can be demonstrated that the model training for MLR and MLP was appropriate, resulting in no overtraining or undertraining. However, in comparison, SVR's model performance was better than MLR and MLP's. Whereas similar to *k*-fold cross-validation findings, the model

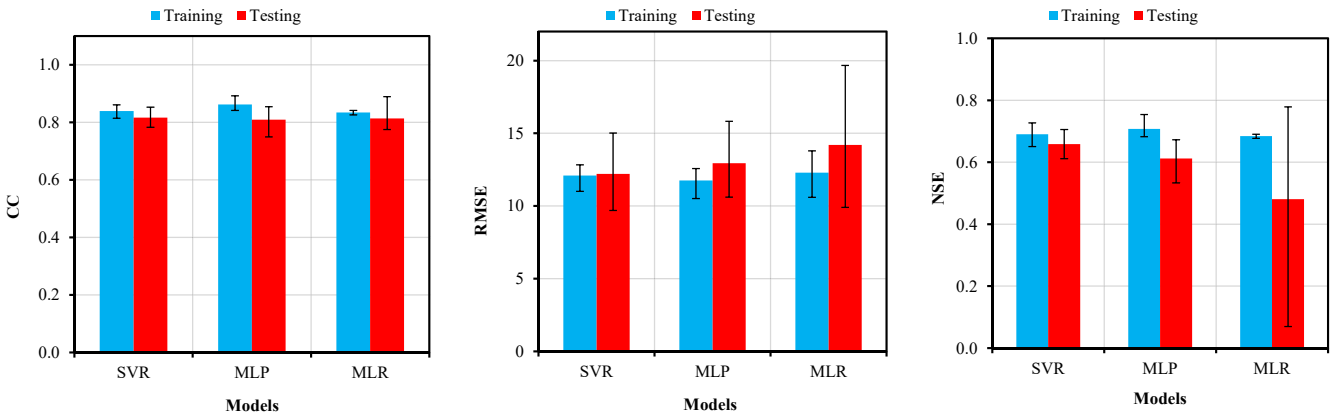


Fig. 10. Mean Correlation Coefficient (CC), Root-Mean-Square Error (RMSE), and Nash–Sutcliffe Efficiency (NSE), along with range bars at the top, show the maximum to the minimum value across different folds for Kala-azar (KA) cases modeling for both Muzaffarpur and Saran districts using Support Vector Regression (SVR), Multiple Linear Regression (MLR), and Multilayer Perceptron (MLP) models.

performance of MLP was observed to be far better than MLR. Hence, on a comparative basis, it can be re-emphasized about the capability of the developed SVR-based model for better simulation of the KA cases in the present study region. Besides, the reason for better performance of SVR may be ascertained by additional advantages being confronted while using RBF kernel SVR instead of traditional SVR. As the former requires fewer hyperparameters, the developed model becomes more accurate under less grid-searching. It has been reported that SVR with RBF kernel becomes comparable with 2-layer neural networks, thereby exploiting the benefits of both traditional SVR and ANN (Achieng, 2019; Du et al., 2021).

4. Discussion

Bihar is the most badly hit state in India, considering the largest number of KA incidences being reported coupled with the presence of the highest number of endemic districts (Goyal et al., 2020; Kumar et al., 2020; Priyamvada et al., 2021; Tesfaye et al., 2017). In conjunction with the spatial anomaly, KA cases are largely reported from the northern region of Bihar (located above the Ganga river) (Fig. 4). Together, north-west and north-east Bihar report more than 80% of the KA cases, followed by south-west Bihar (less than 15%), and the least cases are reported from south-east Bihar. Since the scope of the present study remained surrounded to the most badly hit districts (Muzaffarpur and Saran from the north-west) of Bihar, future research on understanding spatial heterogeneity between best-performing region (south-east) and poor-performing region (north-west) of Bihar may be conducted to further provide research directions and possible solutions to the aforementioned anomaly. However, findings from the present investigations on trends of KA cases, impacts of population density, and influence of changing climate variables in Muzaffarpur and Saran districts also provided deep insights regarding their inter-correlation through the ML-based approach.

Coherently, the study inspected the reasons behind the profound impacts of KA outbreaks in study sites along with their recent fall in terms of numbers. For this, the study undertook population density as a socio-economic variable (discussed here) and climate variables in the changing climate context (discussed in the next paragraph). The results, more or less, identified the rapidly increasing population density, alongside high population heterogeneity, as a primary cause for KA outbreaks. This is because, given that humans contribute to KA transmission dynamics, uneven population distribution (for instance, 90%+ settlement in underdeveloped rural areas as compared to urban) may result in mass congestion. Remarkably, the population density was also identified as a significant variable during the stepwise forward selection and backward elimination process for model input as well as during correlation analysis. Recent studies further reaffirmed the finding (Bhunia et al., 2013; Kumar et al., 2020). Additionally, the interlinkage between population density and KA incidences can also be understood with this analogy: on average, the population density of India is 427 persons/km^2 (https://www.macrotrends.net/countries/IND/india/population-density), and that of Bihar state is 1,102 persons/km^2 (258% higher than the national verage). However, the same was identified as more than 1,400 people/km^2 in the study sites (Fig. 6), which is 328% and 127% higher than the national average and state average, respectively. Congruently, the number of KA cases between 2016 and 2021 in Muzaffarpur and Saran is reported to be 1,830 and 3,372, respectively (Fig. 4), which is not only 11% and 20% of Bihar state but one of the highest in India. In spite of these disturbing figures, the number of KA incidences has been reported to be declining (Fig. 2).

To determine if climate influence existed on a large and falling number of KA cases in Muzaffarpur and Saran districts, statistical tests were conducted for climate variables in the present study site, which included trend variations (Fig. 2), correlation analyses (Figs. 7 and 8), and significant variable selection/elimination tests as was discussed in Sects. 3.1. and 3.2. In general, the findings indicated temperature (T_max, T_min, and T_avg) as the most significant variable, followed by wind speed and then rainfall. Imperatively, the trend test indicated a decreasing trend in temperature between 2016 and 2021 against the increasing trend of rainfall and relative humidity. Since the temperature was recorded as a more significant variable than rainfall and relative humidity, it was logical that the declining warming climate would more negatively influence the rising KA cases than the positive influence of rainfall. Hence, the findings presented in this study argued that the decreasing temperature could be one of the possible reasons for the fall in KA cases. However, beyond climate influence, a fall in the KA cases can also be attributed to the recent developments made in the health infrastructure and the launching of health schemes and programs jointly by the Governments of Bihar and India, specific to combating KA outbreaks [as reported by Kumar & Mishra (2015), Kumar et al. (2020), and Priyamvada et al. (2021)].

Considering the compounding effect of climate variables and population density on KA cases, the study developed an RBF kernel-based SVR model, wherein temperature, wind speed, rainfall, and population density were used as significant input variables. In addition, to validate the model performance, MLR and MLP models were also developed for comparative analysis by conducting a k-fold ($k = 3$) cross-validation analysis (Fig. 9) and by using CC, RMSE, and NSE (refer to Table 2 and Fig. 10). To summarize, SVR-based models alongside MLP and MLR models demonstrated unfailing potential in extracting the KA diseases patterns and outbreaks from merely limited datasets with limited study periods. In addition, the ML-based SVR model results reaffirmed similar

findings from recent literature (Goyal et al., 2020; Kumar et al., 2020; Priyamvada et al., 2021; Tesfaye et al., 2017). The superior performance of the SVR model compared to MLR and MLP models can be attributed to several factors. Firstly, SVR is a nonlinear regression technique that can capture complex relationships between the input variables (climate variables and population density) and the output variable (Kala-azar cases). This flexibility allows SVR to capture non-linear patterns and interactions that may exist in the dataset, providing a more accurate representation of the underlying relationship between the variables. Secondly, SVR utilizes the kernel trick, which maps the input variables into a higher-dimensional feature space. This transformation enables SVR to handle non-linear data better by implicitly performing non-linear transformations without specifying the transformation function. In contrast, MLR assumes a linear relationship between the input and output variables, which may not be sufficient to capture the complexities present in the data. Furthermore, SVR is less susceptible to the influence of outliers compared to MLR and MLP models. SVR aims to find a robust hyperplane that best fits the data while maintaining a maximum margin of separation. This characteristic allows SVR to effectively handle outliers and noise in the dataset, resulting in more robust predictions. Lastly, the performance of the SVR model can be attributed to the appropriate selection of input variables. Through feature selection techniques and domain knowledge, relevant climate variables and population density were identified as significant factors influencing Kala-azar cases. This targeted selection of input variables enhances the model's ability to capture the key drivers of the disease outbreak.

This study's originality stems from its application of machine learning algorithms (SVR, MLR, and MLP) to a critical public health concern (KA outbreaks in India), leveraging the available data to develop predictive models. While previous studies have explored these algorithms for VBD monitoring, the distinctiveness of this research emerges from several key aspects. Firstly, within the last decade, only a handful of investigations have employed these specific algorithms for tracking VBD outbreaks in human populations, and even fewer have focused on Indian conditions (Asadgol et al., 2019; Bhunia & Shit, 2020b). Moreover, the utilization of these algorithms for studying KA outbreaks is relatively unexplored terrain, given its status as an emerging area of research. Furthermore, the unique focus on employing ML-based models, particularly SVR, for epidemiological data analysis and monitoring in the context of KA outbreaks adds to the originality of this work. Notably, the absence of similar studies centered on ML-based KA incidence modeling, especially in endemic regions like India, further highlights the necessity of the research. Importantly, the backdrop of climate change and its impact on VBDs, such as the rising incidences of KA, underscores the significance of results and discussion. This research bridges the gap in understanding the dynamics of KA outbreaks under changing climatic conditions, a field that demands attention due to its real-world implications. The novelty of the study falls under the framework of multimodality and integration of subjective inference and objective inference, as exemplified by Modi et al. (2011). Consequently, these findings can potentially guide new research directions, especially in developing and underdeveloped countries facing the challenges of KA incidence amidst changing climates.

Finally, it is imperative to enumerate a couple of challenges the current investigation witnessed. The study focused on the Muzaffarpur and Saran districts of Bihar, India; the findings may not be directly generalizable to other regions with different ecological and climatic conditions. Besides ML algorithms, Deep Learning (DL) algorithms have achieved remarkable success in handling large-scale datasets and extracting intricate patterns. This study, however, does not consider developing a DL model for KA outbreak modeling. This limitation can be attributed to the fact that DL methods often require substantial data to achieve optimal performance, given to train models effectively and prevent overfitting. The present study encountered the challenge of working with a limited dataset comprising only six years of data on Kala-azar cases. This dataset was provided by the concerned health authority of Bihar, and thus the study had no control over the data availability. Given this constraint, the study thus focused on developing reliable machine-learning models to provide meaningful insights and policy directions based on the available data.

5. Conclusions

This study investigated the implications of climate change on *Visceral leishmaniasis* or Kala-azar (KA) cases, which is a Vector-borne Disease (VBD). The KA has been reported as the second-largest parasitic killer globally, with a 75–95% mortality rate. Published literature has demonstrated the seriousness of rapidly increasing KA cases across 88 endemic countries. More specifically, the origin of KA cases has been marked as South Asia, which reports more than 60% of KA cases annually worldwide. In the context of India, around two-thirds of the KA cases in South Asia are reported from Indian states, from which Bihar alone accounts for more than 50% of the Indian cases. Since past studies hypothesized climate change vulnerabilities as one of the driving causes of the KA outbreak, this study is oriented to focus on climate change implications on KA and population density as one of the socio-economic conditions. In addition, even though epidemic prediction systems for controlling VBD outbreaks have been developed in the past, the applications of the ML approaches remained untouched for modeling and monitoring KA cases. This study, therefore, proposed an RBF kernel-based Support Vector Regression (SVR) model for the most affected endemic regions of Bihar, viz., Muzaffarpur and Saran districts considering KA reportings from January 2016 to July 2021. The study identified dependent variables as KA_cases (Kala-azar cases) and independent variables as T_max (maximum temperature), T_min (minimum temperature), T_avg (average temperature), RH (relative humidity), WS (wind speed), and P_density (population density). Significant variable selection was ascertained using forward selection, backward elimination, and stepwise regression procedures, followed by the k-fold cross-validation technique and the kernel-based SVR algorithm for classification. To compare the results of SVR, two additional models were developed following the same methodology, viz., Multiple Linear Regression (MLR) and Multilayer Perceptron (MLP). The major conclusions drawn from this study are as follows:

- Temperature, wind speed, rainfall, and population density were identified as significant influential climate variables contributing to the KA outbreak.
- A comparative analysis established the superiority of the developed RBF kernel-based SVR model [Correlation Coefficient (CC) = 0.82, Root-Mean-Square Error (RMSE) = 12.20, and Nash–Sutcliffe Efficiency (NSE) = 0.66 during the training phase] as compared to its other counterparts, such as MLR (CC = 0.81, RMSE = 14.20, NSE = 0.48) and MLP (CC = 0.81, RMSE = 12.95, NSE = 0.61) models. The performance of the MLP model was marginally better than MLR.
- Even under limited data availability, the developed RBF kernel-based SVR model could provide acceptable performance and be better than its counterparts. Hence, the developed methodology for the proposed SVR-based model may be useful for the data-scarce region prone to KA cases across the globe.
- The RBF kernel-based SVR model successfully captures the complex relationship between different climatological variables and variations in KA cases. The abilities of the ML-based approaches (such as SVR) in extracting hidden and complex relationships between KA and independent variables are comparatively advantageous considering the deployment of pure climate model-simulated VBD estimates, which demand expertise in climate models. The latter is domain-specific as against the wide interdisciplinary applications of the former.

As stated above, it is also important to note that there is a need for building capacity systems to assist public health authorities in

monitoring the KA spread, learning the climate impacts of outbreaks, and ensuring a timelier health service. The development of ML-based models is one of the recent pathways being experimented within monitoring VBDs. The present study is a proposal for the same considering ML modeling approaches controlling the KA cases in endemic regions of the globe.

CRediT authorship contribution statement

Shubham Kumar: Data curation, Formal analysis, Software. **Aman Srivastava:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Rajib Maity:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The study was partially supported by a project sponsored by Space Application Center (SAC), Indian Space Research Organisation (ISRO), Ahmedabad (Ref. No. IIT/KCSTC/Chair/NEW/P/19-20/09). The authors also acknowledge the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Vector Borne Disease Control Programme (NVBDCP), Patna, Bihar, for providing datasets to accomplish this research. The funding for the Research Scholar (Aman Srivastava) was supported by the Prime Minister's Research Fellowship (PMRF/2401746/21CE91R03) under the Ministry of Education, Government of India. Thanks are due to the Department of Civil Engineering, Indian Institute of Technology (IIT) Kharagpur, for providing laboratory-based support to facilitate modeling.

Funding

No funding was received.

Appendix A. Supplementary data

Supplementary data (comprising Figs. S1 and S2) to this article can be found online at https://doi.org/10.1016/j.eswa.2023.121490.

References

Abbott, B. W., Bishop, K., Zarnetske, J. P., Minaudo, C., Chapin, F. S., Krause, S., et al. (2019). Human domination of the global water cycle absent from depictions and perceptions. *Nature Geoscience, 12*(7), 533–540. https://doi.org/10.1038/s41561-019-0374-y

Achieng, K. O. (2019). Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural networks vs support vector regression models. *Computers & Geosciences, 133*, Article 104320. https://doi.org/10.1016/j.cageo.2019.104320

Ahmed, H., Carter, K. C., & Williams, R. A. (2020). Structure and antiparasitic activity relationship of alkylphosphocholine analogues against Leishmania donovani. *Microorganisms, 8*(8), 1117. https://doi.org/10.3390/microorganisms8081117

Alfred, R., & Obit, J. H. (2021). The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. *Heliyon, 7*(6), e07371. https://doi.org/10.1016/j.heliyon.2021.e07371

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics, 16*(4), 523–531. https://doi.org/10.1080/00401706.1974.10489233

Asadgol, Z., Mohammadi, H., Kermani, M., Badirzadeh, A., & Gholami, M. (2019). The effect of climate change on cholera disease: The road ahead using artificial neural network. *PloS One, 14*(11), e0224813. https://doi.org/10.1371/journal.pone.0224813

Babicki, S., Arndt, D., Marcu, A., Liang, Y., Grant, J. R., Maciejewski, A., & Wishart, D. S. (2016). Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Research*, 44 (W1), W147-W153. doi: 10.1093/nar/gkw419.

Bhunia, G.S., Shit, P.K. (2020a). Introduction of Visceral Leishmaniasis (Kala-azar). In: Spatial mapping and modelling for Kala-azar disease. *SpringerBriefs in Medical Earth Sciences*. Springer, Cham. doi: 10.1007/978-3-030-41227-2_1.

Bhunia, G.S., Shit, P.K. (2020b). Measures and control of Kala-azar. In: Spatial mapping and modelling for Kala-azar disease. *SpringerBriefs in Medical Earth Sciences*. Springer, Cham. doi: 10.1007/978-3-030-41227-2_7.

Bhunia, G. S., Kesari, S., Chatterjee, N., Kumar, V., & Das, P. (2013). Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India. *BMC Infectious Diseases, 13*(1), 1–12. https://doi.org/10.1186/1471-2334-13-64

Boser, B.E., Guyon, I., Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings Fifth annual Workshop on Computational Learning Theory*, Pittsburgh, 144–152. doi: 10.1145/130385.130401.

Bray, M., & Han, D. (2004). Identification of support vector machines for runoff modelling. *Journal of Hydroinformatics, 6*(4), 265–280. https://doi.org/10.2166/hydro.2004.0020

Caminade, C., McIntyre, K. M., & Jones, A. E. (2019). Impact of recent and future climate change on vector-borne diseases. *Annals of the New York Academy of Sciences, 1436*(1), 157–173. https://doi.org/10.1111/nyas.13950

Car, Z., Baressi Šegota, S., Anđelić, N., Lorencin, I., & Mrzljak, V. (2020). Modeling the spread of COVID-19 infection using a multilayer perceptron. *Computational and Mathematical Methods in Medicine*. https://doi.org/10.1155/2020/5714714

Census of Muzaffarpur (2011). Muzaffarpur district: population 2011-2022 data. Retrieved March 17, 2022, from https://www.census2011.co.in/census/district/68-muzaffarpur.html.

Census of Saran (2011). Saran district: population 2011-2022 data. Retrieved March 17, 2022, from https://www.census2011.co.in/census/district/71-saran.html.

CGWB-Muzaffarpur (2013). Ground water information booklet: Muzaffarpur district, Bihar state. *Central Ground Water Board (CGWB)*, Ministry of Water Resources, Government of India, Mid-Eastern Region, Patna. http://cgwb.gov.in/District_Profile/Bihar/Muzaffarpur.pdf.

CGWB- Saran (2013). Ground water information booklet: Saran district, Bihar state. *Central Ground Water Board (CGWB)*, Ministry of Water Resources, Government of India, Mid-Eastern Region, Patna. http://cgwb.gov.in/District_Profile/Bihar/Saran.pdf.

Choy, K. Y., & Chan, C. W. (2003). Modelling of river discharges and rainfall using radial basis function networks based on support vector regression. *International Journal of Systems Science, 34*(14–15), 763–773. https://doi.org/10.1080/00207720310001640241

de Angeli Dutra, D., Salloum, P. M., & Poulin, R. (2023). Vector microbiome: Will global climate change affect vector competence and pathogen transmission? *Parasitology Research, 122*(1), 11–17. https://doi.org/10.1007/s00436-022-07734-x

Dibike, Y. B., Velickov, S., Solomatine, D., & Abbott, M. B. (2001). Model induction with support vector machines: Introduction and applications. *Journal of Computing in Civil Engineering, 15*(3), 208–216. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208)

Ding, X., Liu, J., Yang, F., & Cao, J. (2021). Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute, 358*(18), 10121–10140. https://doi.org/10.1016/j.jfranklin.2021.10.005

Du, B., Lund, P. D., Wang, J., Kolhe, M., & Hu, E. (2021). Comparative study of modelling the thermal efficiency of a novel straight through evacuated tube collector with MLR, SVR, BP and RBF methods. *Sustainable Energy Technologies and Assessments, 44*, Article 101029. https://doi.org/10.1016/j.seta.2021.101029

Elbeltagi, A., Raza, A., Hu, Y., Al-Ansari, N., Kushwaha, N. L., et al. (2022). Data intelligence and hybrid metaheuristic algorithms-based estimation of reference evapotranspiration. *Applied Water Science, 12*(7), 152. https://doi.org/10.1007/s13201-022-01667-7

Elbeltagi, A., Srivastava, A., Kushwaha, N. L., Juhász, C., Tamás, J., & Nagy, A. (2023a). Meteorological data fusion approach for modeling crop water productivity based on ensemble machine learning. *Water, 15*(1), 30. https://doi.org/10.3390/w15010030

Elbeltagi, A., Srivastava, A., Al-Saeedi, A. H., Raza, A., Abd-Elaty, I., & El-Rawy, M. (2023b). Forecasting long-series daily reference evapotranspiration based on best subset regression and machine learning in Egypt. *Water, 15*(6), 1149. https://doi.org/10.3390/w15061149

Fouque, F., & Reeder, J. C. (2019). Impact of past and on-going changes on climate and weather on vector-borne diseases transmission: A look at the evidence. *Infectious Diseases of Poverty, 8*(1), 1–9. https://doi.org/10.1186/s40249-019-0565-1

Franklinos, L. H., Jones, K. E., Redding, D. W., & Abubakar, I. (2019). The effect of global change on mosquito-borne disease. *The Lancet Infectious Diseases*, 19(9), e302-e312. doi: 10.1016/S1473-3099(19)30161-6.

Gil, Z., Martinez-Sotillo, N., Pinto-Martinez, A., Mejias, F., Martinez, J. C., Galindo, I., et al. (2020). SQ109 inhibits proliferation of Leishmania donovani by disruption of intracellular Ca2+ homeostasis, collapsing the mitochondrial electrochemical potential ($\Delta\Psi$m) and affecting acidocalcisomes. *Parasitology Research, 119*(2), 649–657. https://doi.org/10.1007/s00436-019-06560-y

Gopi, A. P., Jyothi, R. N. S., Narayana, V. L., & Sandeep, K. S. (2023). Classification of tweets data based on polarity using improved RBF kernel of SVM. *International Journal of Information Technology, 15*(2), 965–980. https://doi.org/10.1007/s41870-019-00409-4

Goyal, V., Das, V. N. R., Singh, S. N., Singh, R. S., Pandey, K., Verma, N., et al. (2020). Long-term incidence of relapse and post-kala-azar dermal leishmaniasis after three

different visceral leishmaniasis treatment regimens in Bihar, India. *PLoS Neglected Tropical Diseases, 14*(7), e0008429. https://doi.org/10.1371/journal.pntd.0008429
Grégoire, G. (2014). Multiple linear regression. *European Astronomical Society Publications Series, 66*, 45–72. https://doi.org/10.1197/j.aem.2003.09.006
Guyon, I., Boser, B., & Vapnik, V. (1992). Automatic capacity tuning of very large VC-dimension classifiers. *Advances in Neural Information Processing Systems*, 5. Retrieved April 18, 2022, from https://proceedings.neurips.cc/paper/1992/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.
Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 1733-1762. https://www.jstor.org/stable/23566492.
Hekmatmanesh, A., Wu, H., Jamaloo, F., Li, M., & Handroos, H. (2020). A combination of CSP-based method with soft margin SVM classifier and generalized RBF kernel for imagery-based brain computer interface applications. *Multimedia Tools and Applications, 79*, 17521–17549. https://doi.org/10.1007/s11042-020-08675-2
IPCC (2021). Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the *Intergovernmental Panel on Climate Change*. https://www.ipcc.ch/report/ar6/wg1/.
Jimenez, F., Palma, J., Sanchez, G., Marin, D., Palacios, M. F., & López, M. L. (2020). Feature selection based multivariate time series forecasting: An application to antibiotic resistance outbreaks prediction. *Artificial Intelligence in Medicine, 104*, Article 101818. https://doi.org/10.1016/j.artmed.2020.101818
Joshi, A., & Miller, C. (2021). Review of machine learning techniques for mosquito control in urban environments. *Ecological Informatics, 61*, Article 101241. https://doi.org/10.1016/j.ecoinf.2021.101241
Karunaweera, N. D., & Ferreira, M. U. (2018). Leishmaniasis: Current challenges and prospects for elimination with special focus on the South Asian region. *Parasitology, 145*(4), 425–429. https://doi.org/10.1017/S0031182018000471
Kumar, G., & Mishra, R. (2015). *Tackling the Kala-Azar Memance in Bihar*. Prabhat Prakashan.
Kumar, S., Roshni, T., Kahya, E., & Ghorbani, M. A. (2020). Climate change projections of rainfall and its impact on the cropland suitability for rice and wheat crops in the Sone river command, Bihar. *Theoretical and Applied Climatology, 142*(1), 433–451. https://doi.org/10.1007/s00704-020-03319-9
Kumar, V., Mandal, R., Das, S., Kesari, S., Dinesh, D. S., Pandey, K., et al. (2020). Kala-azar elimination in a highly-endemic district of Bihar, India: A success story. *PLoS Neglected Tropical Diseases, 14*(5), Article e0008254. https://doi.org/10.1371/journal.pntd.0008254
Kumar, V., Siddiqui, N. A., Pollington, T. M., Mandal, R., Das, S., Kesari, S., et al. (2022). Impact of intensified control on visceral leishmaniasis in a highly-endemic district of Bihar, India: An interrupted time series analysis. *Epidemics, 39*, Article 100562. https://doi.org/10.1016/j.epidem.2022.100562
Lindsey, C., & Sheather, S. (2010). Variable selection in linear regression. *The Stata Journal*, 10(4), 650-669. doi: 10.1177%2F1536867X1101000407.
Liong, S. Y., & Sivapragasam, C. (2002). Flood stage forecasting with support vector machines 1. *Journal of the American Water Resources Association, 38*(1), 173–186. https://doi.org/10.1111/j.1752-1688.2002.tb01544.x
Mahajan, R., Owen, S., Kumar, S., Kazmi, S., Pandey, K., Verma, N., et al. (2023). Prevalence of asymptomatic leishmania infection in people living with HIV and progression to symptomatic visceral leishmaniasis in Bihar, India. *International Journal of Infectious Diseases, 130*, S21. https://doi.org/10.1016/j.ijid.2023.04.051
Maity, R., Bhagwat, P. P., & Bhatnagar, A. (2010). Potential of support vector regression for prediction of monthly streamflow using endogenous property. *Hydrological Processes: An International Journal, 24*(7), 917–923. https://doi.org/10.1002/hyp.7535
Modi, S., Lin, Y., Cheng, L., Yang, G., Liu, L., & Zhang, W. J. (2011). A socially inspired framework for human state inference using expert opinion integration. *IEEE/ASME Transactions on Mechatronics, 16*(5), 874–878. https://doi.org/10.1109/TMECH.2011.2161094
Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology, 10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5), e262-e273. doi: 10.1016/S1470-2045(19)30149-4.
Nguyen, H., Bui, X. N., Choi, Y., Lee, C. W., & Armaghani, D. J. (2021). A novel combination of whale optimization algorithm and support vector machine with different kernel functions for prediction of blasting-induced fly-rock in quarry mines. *Natural Resources Research, 30*, 191–207. https://doi.org/10.1007/s11053-020-09710-7
Okoro, O. J., Deme, G. G., Okoye, C. O., Eze, S. C., Odii, E. C., Gbadegesin, J. T., et al. (2023). Understanding key vectors and vector-borne diseases associated with freshwater ecosystem across Africa: Implications for public health. *Science of The Total Environment, 862*, Article 160732. https://doi.org/10.1016/j.scitotenv.2022.160732
Pal, K., & Patel, B. V. (2020). Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. *Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020*, 83–87. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00016
Pande, C. B., Al-Ansari, N., Kushwaha, N. L., Srivastava, A., Noor, R., Kumar, M., et al. (2022). Forecasting of SPI and meteorological drought based on the artificial neural network and M5P model tree. *Land, 11*(11), 2040. https://doi.org/10.3390/land11112040
Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 187*, 253–318. https://doi.org/10.1098/rsta.1896.0007
Pradhan, S. P., Joshi, P., Poudel, P., Ghimire, A., Chhetri, S., Maharjan, J., et al. (2022). Long-term assessment of water quality of Kathmandu University Drinking Water Supply Centre. *Nepal. Sustainable Water Resources Management, 8*(2), 41. https://doi.org/10.1007/s40899-022-00636-x
Priyamvada, K., Bindroo, J., Sharma, M. P., Chapman, L. A., Dubey, P., Mahapatra, T., et al. (2021). Visceral leishmaniasis outbreaks in Bihar: Community-level investigations in the context of elimination of kala-azar as a public health problem. *Parasites & Vectors, 14*(1), 1–11. https://doi.org/10.1186/s13071-020-04551-y
Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database Systems, 5*, 532–538. https://doi.org/10.1007/978-1-4899-7993-3_565-2
Rocklöv, J., & Dubrow, R. (2020). Climate change: An enduring challenge for vector-borne disease prevention and control. *Nature Immunology, 21*(5), 479–483. https://doi.org/10.1038/s41590-020-0648-y
Rosenblatt, F. (1961). *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Cornell Aeronautical Lab Inc Buffalo NY. Retrieved April 18, 2022, from https://apps.dtic.mil/sti/pdfs/AD0256582.pdf.
Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x
Scavuzzo, J. M., Trucco, F., Espinosa, M., Tauro, C. B., Abril, M., Scavuzzo, C. M., et al. (2018). Modeling Dengue vector population using remotely sensed data and machine learning. *Acta Tropica, 185*, 167–175. https://doi.org/10.1016/j.actatropica.2018.05.003
Srivastava, A., Maity, R., & Desai, V. R. (2022). Assessing global-scale synergy between adaptation, mitigation, and sustainable development for projected climate change. In U. Chatterjee, A. O. Akanwa, S. Kumar, S. K. Singh, & A. Dutta Roy (Eds.), *Ecological footprints of climate change*. Cham: Springer Climate. Springer. https://doi.org/10.1007/978-3-031-15501-7_2.
Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications, 134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028
Tapak, L., Hamidi, O., Fathian, M., & Karami, M. (2019). Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC Research Notes, 12*(1), 1–6. https://doi.org/10.1186/s13104-019-4393-y
Taud, H., Mas, J. (2018). Multilayer perceptron (MLP). In: Camacho Olmedo, M., Paegelow, M., Mas, JF., Escobar, F. (Eds.) Geomatic approaches for modeling land change scenarios. *Lecture Notes in Geoinformation and Cartography*. Springer, Cham. doi: 10.1007/978-3-319-60801-3_27.
Tesfaye, K., Aggarwal, P. K., Mequanint, F., Shirsath, P. B., Stirling, C. M., Khatri-Chhetri, A., et al. (2017). Climate variability and change in Bihar, India: Challenges and opportunities for sustainable crop production. *Sustainability, 9*(11), 1998. https://doi.org/10.3390/su9111998
Tranmer, M., & Elliot, M. (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research*, 5(5), 1-5. Retrieved April 18, 2022, from http://humme dia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf.
Vapnik, V. N. (1998). *Statistical learning theory* (p. 1). New York: John Wiley & Sons. Inc..
Vapnik, V.N. (2000). The nature of statistical learning theory. *Statistics for Engineering and Information Science*. Springer, New York, NY. doi: 10.1007/978-1-4757-3264-1.
Wilcox, B. A., Echaubard, P., de Garine-Wichatitsky, M., & Ramirez, B. (2019). Vector-borne disease and climate change adaptation in African dryland social-ecological systems. *Infectious Diseases of Poverty, 8*(1), 1–12. https://doi.org/10.1186/s40249-019-0539-3
Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82. https://www.int-res.com/articles/cr2005/30/c030p079.pdf.
Wilson, A. L., Courtenay, O., Kelly-Hope, L. A., Scott, T. W., Takken, W., Torr, S. J., et al. (2020). The importance of vector control for the control and elimination of vector-borne diseases. *PLoS Neglected Tropical Diseases, 14*(1), e0007831. https://doi.org/10.1371/journal.pntd.0007831
Xu, L., & Zhang, W. J. (2001). Comparison of different methods for variable selection. *Analytica Chimica Acta, 446*(1–2), 475–481. https://doi.org/10.1016/S0003-2670(01)01271-5
Yadav, P., Azam, M., Ramesh, V., & Singh, R. (2023). Unusual observations in Leishmaniasis—An overview. *Pathogens, 12*(2), 297. https://doi.org/10.3390/pathogens12020297