Corpus Intellectual
ISSN PRINT 2811-3187 ONLINE 2811-3209 Volume 4 NO 3 2025

Leveraging Machine Learning Techniques for Variable Selection in Modelling Malaria Transmission

Lead Author

Oluyemi Adewole Okunlola.

Affiliation:

Department of Mathematics and Statistics, Redeemer's University Ede, Osun State, Nigeria.

Abstract

Disease modelling is no exception to the widespread acceptance of artificial intelligence across many fields, particularly in cases where a group of highly correlated predictors is linked to the disease's outcome. In order to avoid misleading regression coefficients and inflated standard errors in modelling, multicollinearity must still be absent. With data consisting of predictors that are inherently associated, this study focuses on variable selection in modelling the spread of malaria. Thirteen predictors were analysed, including topography, livestock indices, environmental, and control measure variables. The study addressed multicollinearity in an effort to increase the model's predictive capacity by utilizing machine learning components of artificial intelligence. In particular, regularized algorithms like ridge, elastic net and least absolute and shrinkage selection operator (LASSO) were taken into consideration. A Poisson-based random forest was employed as a comparison tool. There was a small difference between the three regularization method variants under comparison based on the mean square error and R-square performance measurements. LASSO outperformed the other two techniques as evidenced by the lowest mean square error value (LASSO = 2.017042, ridge = 2.022117, elastic net = 2.023353). Though the number of important predictors chosen for malaria transmission was precisely the same as that of LASSO, the mean square error of the ensemble Poisson-based random forest (MSE = 0.006167) was much lower than that of the regularisation techniques.

The random forest value was twice as high as the regularization algorithm value in terms of R-square. Artificial intelligence is a crucial tool to solve the issues posed by big data as it continues to evolve, especially in the area of disease modelling.

Keywords: Multicollinearity, Regularisation, Artificial intelligence, Disease modelling, Random Forest, Poisson

Co-authors: Dorcas Modupe Okewole, Department of Mathematics and Statistics, Redeemer's University. **Idowu Peter Adewumi**, Department of Health Information Management, University of Medical Sciences, Ondo, Ondo State, Nigeria. **Adewale Folaranmi Lukman**, Department of Mathematics and Statistics, University of North Dakota, USA.

Introduction

Malaria is a severe public health concern, exposing about half of the global population to the risk of infection, with a disproportionate effect on tropical and subtropical areas where the conditions are favourable for its spread (Centers for Disease Control and Prevention (CDC), 2024). In 2022, the global malaria cases and related deaths were around 249 million and 608 thousand, respectively, and the African continent reported 94% of cases of malaria and 80% of related mortalities (World Health Organisation (WHO), 2022). The most vulnerable population in this region was children under the age of five (Shekarau et al., 2024). This alarming report calls for comprehensive initiatives to detect and prevent malaria transmission among regions that are disproportionately impacted (Awasthi et al., 2024).While malaria transmission is influenced by a complex interaction of biological, social, and environmental factors, to capture the inherent complexities associated with many correlated variables, the diverse character of malaria transmission needs advanced analytical techniques that go beyond traditional regression models (Shretta et al., 2017).

Multicollinearity is a major issue in malaria research, causing standard epidemiological models to fail when dealing with datasets containing strongly linked components(Shrestha, 2020). Multicollinearity can complicate the underlying impact of individual variables by increasing standard errors and causing unstable regression coefficients, making the results difficult to understand and rely on (Ochieng, 2024). Because multicollinearity reduces models' ability to offer precise forecasts and relevant insights, it is especially important in malaria modelling, where variables range from socioeconomic indices to

meteorological conditions. Addressing this difficulty is necessary for improving the interpretability and effectiveness of malaria research models (Amadi & Erandi, 2024).

Effective variable selection is critical for improving malaria model prediction performance and ensuring that the results are relevant to public health operations and this is important for developing a more efficient and precise model since it helps discover the most important predictors from a huge pool of highly correlated variables(Aheto et al., 2021a). While previous research has demonstrated that variable selection can reduce the negative effects of multicollinearity on model stability and interpretability, eventually boosting the model's ability to capture critical malaria transmission factors, variable selection is more than just a statistical exercise; it is also an important stage in developing models that can successfully lead public health actions focused at malaria reduction(Chan et al., 2022; Kyriazos & Poga, 2023).

The failure of previous malaria modelling tools to appropriately address multicollinearity across variables is a significant problem, and when predictors are highly correlated, as is common in studies including biological and environmental variables, approaches like simple least squares regression sometimes generate incorrect results(Stanley et al., 2019; Vatcheva & Lee, 2016). This has motivated academics to investigate machine learning techniques as a more efficient method of variable selection in high-dimensional datasets on malaria transmission and malaria research can produce more accurate and reliable modelling findings due to machine learning techniques' increased ability to manage huge datasets with various interrelated parts (Adamua & Singh, 2021; O. Khan et al., 2024).

While machine learning has proven to be an effective tool for dealing with multicollinearity and high-dimensional data in disease modelling, regularisation techniques such as Ridge regression, Elastic Net, and Least Absolute Shrinkage and Selection Operator (LASSO) have been shown to be effective in variable selection by introducing penalties that shrink coefficients, reducing the effect of multicollinearity on model performance (Aheto et al., 2021a; Shrestha, 2020). For example, LASSO eliminates non-informative predictors from the model by imposing an L1 penalty, which causes some coefficients to become zero. This feature is notably valuable in malaria modelling because it allows for the identification of critical socioeconomic and environmental elements that influence malaria transmission without being confused by the effects of multicollinearity(Guo et al., 2015).

Elastic Net is a useful tool when numerous variables are highly correlated since it combines the advantages of Ridge regression and LASSO, this strategy has showed promise in malaria research, where environmental parameters such as temperature, rainfall, and vegetation indices are commonly associated(Aheto et al., 2021a). Elastic Net is a promising method for infectious disease modelling because it reduces overfitting and increases model resilience by incorporating both L1 and L2 penalties (Z. Li et al., 2020). Ridge regression, on the other hand, only applies an L2 penalty, resulting in smaller regression coefficients, thereby stabilising the model and makes it suitable for scenarios in which all predictors are assumed to be relatively relevant(Aheto et al., 2021a).

These machine learning algorithms increase forecast accuracy and help researchers better understand the factors that lead to malaria transmission. Okunlola et al., (2021) demonstrated how regularised regression approaches can be used to choose relevant environmental characteristics, hence lowering the computing complexity of malaria models. These strategies promote enhanced model interpretability and data efficiency, both of which are critical for generating targeted interventions in high-risk areas(Lucas et al., 2022).The application of advanced machine learning techniques in malaria research is justified by the inherent complexity of malaria transmission and the enormous number of linked factors(O. Khan et al., 2024).

While traditional statistical approaches are valuable, they are insufficient to handle high-dimensional data with connected variables(Filzmoser & Nordhausen, 2021). Machine learning methods, particularly regularisation and ensemble approaches, provide a more rigorous framework for discovering the primary elements impacting malaria spread(Mujahid et al., 2024).Using these advanced methodologies, reliable and interpretable models can be constructed, providing valuable information for public health initiatives. According to Wiemken & Kelley (2019), machine learning approaches are becoming increasingly essential in epidemiology due to their capacity to solve issues such as variable significance and multicollinearity in large datasets. These advancements demonstrate how machine learning may increase the precision and understandability of malaria models, resulting in better disease management methods in the long term(Kino et al., 2021).

Therefore, given the continuous issues associated with malaria transmission modelling, this study employs machine learning techniques to address the multicollinearity problem in high-

dimensional malaria data, thus establishing the optimal approach for variable selection in malaria transmission modelling by comparing the performance of regularised regression approaches (particularly LASSO, Ridge, and Elastic Net). The findings of this work will considerably enhance the field of public health by shedding light on the most effective pointers of malaria transmission and illustrating how machine learning may increase the predictive accuracy of malaria models.

Variable selection in malaria modelling or prediction using machine learning

Ridge Regression in Malaria Modeling

Ridge regression, an L2 regularisation technique, is commonly employed in malaria modelling to manage multicollinearity and overfitting, especially when predictors like socioeconomic and climatic factors are heavily connected(Schreiber-Gregory, 2018). While ridge increases prediction accuracy and model stability by penalising big coefficients while retaining predictors, it has been efficient in forecasting complicated interactions between environmental parameters like as temperature and rainfall that affect malaria transmission in disease research(Aheto et al., 2021a; Obadina et al., 2021). By retaining the predictive power of these associated characteristics, the technique strengthens the model and minimises its susceptibility to overfitting(Ying, 2019).

Ridge regression's inability to pick variables is a serious constraint, particularly in malaria modelling, where interpretability is critical. Ridge typically retains predictors with little effect on the outcome by reducing coefficients rather than removing unnecessary variables, making it more difficult to identify the important elements influencing malaria spread(Obadina et al., 2021). While ridge is less appropriate for investigations designed at determining the individual contributions of each component due to its tradeoff between interpretability and accuracy, its retention of all parameters may diminish the clarity required in malaria research for designing personalised medications, despite the fact that it is extremely effective at prediction tasks(Usman et al., 2022). Ridge regression is thus less ideal for research seeking to uncover major drivers of malaria transmission, despite its usefulness for dealing with multicollinearity and improving model stability(Cule & De Iorio, 2013).

LASSO Regression in Malaria Prediction

LASSO regression, a regularisation technique, does both variable selection and shrinkage by imposing an L1 penalty on the coefficients, effectively eliminating non-contributory predictors from the model by causing some coefficients to be zero(Guo et al., 2015). LASSO is very beneficial for simplifying malaria prediction models because it reduces the dimensionality of high-dimensional datasets by removing extraneous variables while retaining the ones that are most significant for assessing malaria transmission(Yang & Wen, 2018). Due to this feature, LASSO has been extremely beneficial in simulating malaria, where a vast variety of socioeconomic, environmental, and health-related factors are often closely connected(Aheto et al., 2021b). Manguin et al. (2018) and Opiyo et al. (2021) demonstrate how LASSO can detect essential variables such as temperature, rainfall, and vector control actions while removing extraneous or redundant features, hence improving model interpretability(Agrawal, 2023).

One important disadvantage of LASSO is its ability to under-select variables in instances where predictor variables are closely connected. In some cases, LASSO may choose one predictor at random while disregarding others that are equally essential, ignoring intricate inter-variable interactions(Yazdi et al., 2021). Furthermore, the approach may be sensitive to the tuning parameter that governs the intensity of the penalty; hence, the outcome will be determined by the parameter used(Yazdi et al., 2021). Despite these challenges, LASSO remains useful for simulating malaria, particularly in research focused at discovering and assessing the most relevant elements impacting malaria transmission(Muthukrishnan & Rohini, 2017). Due to its potential to improve model clarity and minimise complexity, it is an effective tool for developing tailored, data-driven malaria intervention programmes(Aheto et al., 2021).

Elastic Net in Malaria Modeling

Elastic Net is a regularisation method that combines the benefits of Ridge and LASSO regression to solve multicollinearity and variable selection in high-dimensional datasets. It is a hybrid technique that achieves variable selection while simultaneously increasing coefficient shrinkage, making it useful for dealing with correlated predictors(Usman et al., 2022). This combination allows Elastic Net to preserve the benefits of LASSO in discovering critical variables while also utilising Ridge's stability in dealing with multicollinearity(Aheto et al., 2021). In the field of malaria modelling, where datasets typically contain highly correlated environmental, socioeconomic, and

demographic variables, Elastic Net has proven useful in identifying key predictors while maintaining model robustness(Obasohan et al., 2021). Bayoh et al. (2019) employed Elastic Net to mimic malaria transmission patterns, emphasising its capacity to manage complex interactions among several variables without overfitting(Gimba & Bala, 2017).

Elastic Net provides substantial advantages in regulating related predictors and enhancing model stability, but it also has certain downsides. Optimising the two regularisation parameters (α for L1 and λ for L2) is computationally intensive and requires cross-validation(Gimba & Bala, 2017). The model's success is largely determined by the balance of LASSO and Ridge penalties, and incorrect tuning may result in suboptimal variable selection or predictive performance(De Mol et al., 2009). Despite these limitations, Elastic Net is well-suited for malaria modelling, especially in datasets with complicated correlation structures, making it a promising tool for finding important malaria transmission factors and guiding targeted intervention strategies (Rauschenberger et al., 2021).

Application of Machine Learning Techniques in Malaria Research

Machine learning (ML) approaches in malaria research represent an innovative approach to understanding and forecasting malaria spread and recent improvements demonstrate how machine learning models may combine a variety of data sets, including demographic, socioeconomic, and meteorological information, to produce more reliable and accurate malaria prediction models(Mujahid et al., 2024). For example, LASSOand other regularisation approaches have been successfully applied to analyse big datasets, reducing data dimensionality and discovering relevant predictors of malaria occurrence(Yazdi et al., 2021). In terms of predictive power and model stability, these approaches outperformed established statistical techniques such as linear regression (Muthukrishnan & Rohini, 2017).Machine learning algorithms such as Random Forest and support vector machines have been demonstrated to predict malaria epidemics in The Gambia based on environmental characteristics such as temperature, humidity, and rainfall patterns (O. Khan et al., 2024).

These models highlighted the significance of adding real-time environmental data into malaria management approaches, as well as successfully predicting malaria cases. Beyond predictive modelling, machine learning approaches provide excellent tools for investigating intricate interactions between several factors, particularly the dynamics of malaria transmission, which are influenced by a variety of

characteristics(Adamua & Singh, 2021; Mujahid et al., 2024). While machine learning algorithms can detect nuanced interactions between variables, such as the synergy between socioeconomic characteristics and climate conditions, standard statistical methods frequently fail to account for these nonlinear correlations (Gaye et al., 2021). Even when factors are highly linked, research works utilising Elastic Net have demonstrated that this method may identify the most relevant predictors, providing information that can be used to guide personalised malaria treatments (Aheto et al., 2021a). By uncovering these underlying patterns, machine learning has helped to improve malaria forecasting models and facilitate more successful public health campaigns, particularly in high-burden countries(O. Khan et al., 2024).

Challenges and Future Directions

Even though machine learning holds enormous potential for simulating malaria, some difficulties must be addressed before its full potential is reached. The selection of appropriate algorithms is critical because, depending on the structure and quality of the data, different machine learning models might generate vastly different outcomes(Olushola et al., 2023). Incomplete, biassed, or insufficient data can result in erroneous predictions, therefore the quality and quantity of available data are critical factors(Nugroho, 2023). Even though machine learning excels at dealing with massive datasets, the interpretability of these models remains a key issue(Nilashi et al., 2023). This is especially true for stakeholders such as public health professionals, who rely on precise, usable information to make decisions. Despite their strength, complicated models such as Elastic Net may not provide the necessary transparency in policy-making settings, preventing their continued use in malaria prevention activities (De Mol et al., 2009).

Furthermore, modifying hyperparameters, which is critical for maximising model performance, remains a time-consuming and costly computing procedure and when processing capacity is constrained, using machine learning may become less practicable(Yu & Zhu, 2020).Another problem is integrating machine learning models with conventional epidemiology approaches. Although machine learningcan provide useful insights into the dynamics of malaria transmission, when integrated with more traditional methods such as statistical epidemiology, it may result in more trustworthy models that provide a better understanding of the disease's causes(Li & Abdallah, 2022). To ensure that machine learning models are both practical and scientifically sound, data scientists, epidemiologists, and public health specialists must collaborate(Raiaan et al., 2024).

In the future, combining advanced machine learning techniques like as regularisation and Poisson-based random forests could be a feasible solution to these issues. Malaria treatments can be more targeted and efficient if machine learning improves prediction accuracy and identifies major transmission drivers such as socioeconomic and climatic variables(Orimadegun & Ilesanmi, 2015). However, continued research and improvement are required to ensure that these techniques are the best, most interpretable, and most adaptable for implementation in actual malaria control efforts.
Model and Estimation

Let $y_i$be a response variable having a Poisson distribution, and $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip})^T$, *i*=1,2…,*n,* represent the p-dimensional vector of predictors of the *i-th* observation. The Poisson regression model is formulated as follows:

$$P(Y_i = y_i|x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, y_i = 0,1,2,\dots,\lambda_i > 0, \quad (2.1)$$

where λ_i denotes both the mean and variance, $E(Y_i) = V(Y_i) = \mu_i$. To incorporate covariates into the model, the model parameter λ_i is modeled using the log-link function:

$$\log(\mu_i) = x_i^T\beta, \quad (2.2)$$

where $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)^T$is a vector of unknown regression coefficients. The log-likelihood function for this model is given by:

$$\ell(\beta) = \sum_{i=1}^{n}[y_i \log(\mu_i) - \mu_i - log(y_i!)]. \quad (2.3)$$

Ridge Estimation

In the presence of multicollinearity, the maximum likelihood estimates of the coefficients can be unstable. Ridge regression by Hoerl and Kennard (1970) offers a remedy by introducing an L_2 -norm penalty, which shrinks the regression coefficients towards zero:

$$\hat{\beta}_{ridge} = arg\min_{\beta}\left\{-\ell(\beta|y) + \lambda \sum_{j=1}^{p}\beta_j^2\right\}, \quad (2.4)$$

where $\gamma > 0$ is a tuning parameter controlling the strength of the penalty. This regularization helps to reduce variance in the estimated coefficients by shrinking their values. Ridge regression, however, shrinks

all coefficients but does not perform variable selection, meaning all predictors remain in the model, though with reduced magnitudes. Its strength lies in improving stability and interpretability in cases of multicollinearity.

Lasso Estimation

The Least Absolute Shrinkage and Selection Operator (Lasso) method, introduced by Tibshirani (1996), provides both shrinkage and variable selection by applying an $L_1 - \mathbf{norm}$ penalty to the regression coefficients:

$$\hat{\beta}_{lasso} = arg\min_{\beta}\left\{-\ell(\beta|y) + \lambda\sum_{j=1}^{p}|\beta_j|\right\}. \quad (2.5)$$

The L_1-penalty can shrink some coefficients exactly to zero, thereby excluding irrelevant predictors from the model. This feature makes Lasso particularly useful in high-dimensional datasets, where many predictors might be redundant or irrelevant. By performing variable selection automatically, Lasso enhances the model's interpretability and simplifies the final predictive model, focusing on the most important predictors.

Elastic Net Estimation

Elastic Net, proposed by Zou and Hastie (2005), is a ridge and Lasso regression hybrid. It combines the L_1 and L_2-norm penalties to handle situations where predictors are highly correlated. The Elastic Net estimator is defined as:

$$\hat{\beta}_{elastic-net} = arg\min_{\beta}\left\{-\ell(\beta|y) + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right\}. \quad (2.6)$$

where λ_1 and λ_2 are non-negative tuning parameters. The L_1-penalty encourages sparsity, while the L_2-penalty stabilizes the solution by shrinking the coefficients. Elastic Net is especially useful when there are groups of highly correlated variables, as it can select or drop such groups collectively.

Lasso and Elastic Net play crucial roles in variable selection, a process essential for constructing more interpretable and parsimonious models, especially when the number of predictors is large. Lasso achieves sparsity by selecting individual predictors, which is vital in high-dimensional settings where many variables may be irrelevant.

Elastic Net, by blending Lasso's sparsity with Ridge's grouping effect, ensures that not only are key predictors selected, but highly correlated groups of predictors are either retained or eliminated together, leading to more stable and interpretable models. This ability to automatically select variables while controlling for multicollinearity makes these penalized regression techniques indispensable for handling complex, high-dimensional datasets efficiently.

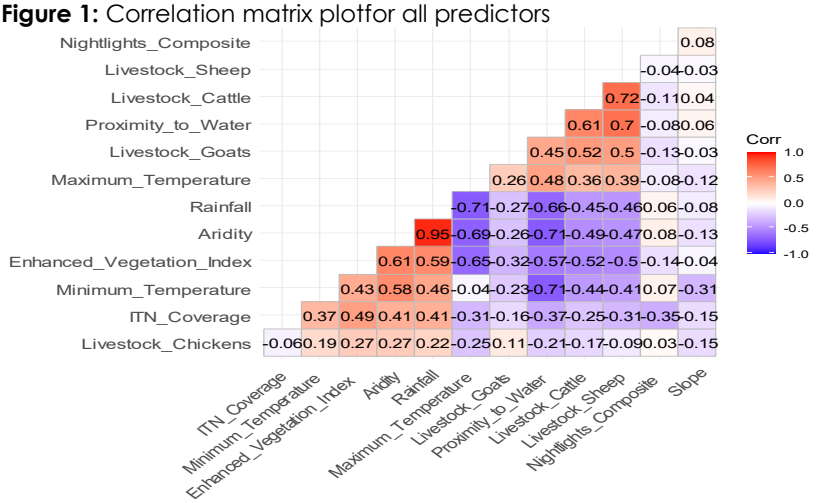
Data and Statistical Analysis

The data used in the study were sourced from the recent nationally representative demographic and health surveys of Nigeria conducted in 2018. The extracted features include environmental factors (aridity, rainfall, enhanced vegetation index, minimum and maximum temperature), livestock indices (chickens, goats, cattle and sheep) topographic variable (slope) and socio-economic variables which include insecticide bed-net coverage(ITN), proximity to water and night light time while the target variable is malaria incidence. All the variables were in continuous form and this necessitated the standardization of all the features prior to data analysis. To avoid over-fitting, the data were divided into training and test set in the ratio of 75 to 25 and k folds cross validation was adopted in the model building. As against the division of the data into the training and test where the model is built with training set and evaluated on the test set just one, k folds cross validation is a resampling procedure with a single parameter k that refers to the number of groups that a given data sample is to be split into. In this case, we set $k = 10$implying the training set is randomly divided into ten parts. Each subset is considered as the test set and the remaining subsets are used to train the model. Lamda (λ) is the penalty term in regulation algorithm and its value need to be determined through cross validation. 10 fold cross validation was used to determine λ value for each of the three algorithms. α is fixed between 0 and 1. It is zero for ridge, one for LASSO and between 0 and 1 for elastic net.

Result

The correlation matrix plot presented in Figure 1 revealed that some pairs of the predictors are correlated. For instance, rainfall and aridity; rainfall and maximum temperature; livestock sheep and proximity to water as well as livestock cattle with correlation coefficients $r = 0.95, -0.71, 0.70$ and 0.72, respectively. The high correlation coefficient value of predictors' variable pairs is a signal of multicollinearity.

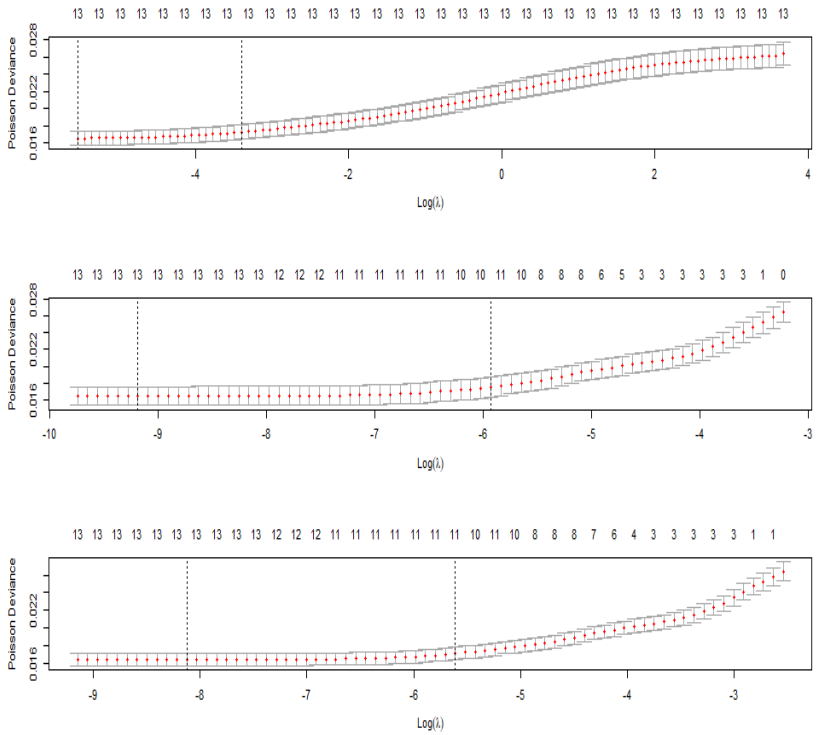
Figure 1: Correlation matrix plotfor all predictors



Source: Authors' construct from the data

Feature selection was undertaken using the regulation algorithms (ridge, lasso and elastic net). The penalty parameter (λ) of the technique was determined with 10 folds cross validation. This value was found to be 0.03349, 0.002654 and 0.003659 for the ridge, lasso and elastic net, respectively. Figure 2 depicts the cross-validation curve (red dotted line), with Poisson deviance, log of penalty parameter and number of non-zero predictors (upper horizontal axis). The vertical dotted lines along the λ sequence represent two special values. The value of λ that yields the least mean cross-validated error is lambda. min (first vertical dotted line), and the value of λ that yields the most regularized model such that the cross-validated error is within one standard error of the minimum is lambda.1se (second dotted vertical line).Rows 1, 2, and 3 of Figure 1 correspond to the ridge, lasso, and elastic net cross-validation plot for the log of lambda. It is important to note from the figure that as λ, the number of non-zero coefficients increase. For ridge, no coefficient was shrunk to zero, and the thirteen predictors were retained. In the lasso and elastic net, 10 and 11 predictors, respectively, were retained while others were shrunk to zero due to collinearity.

Figure 2: Cross-validation plot for ridge, lasso and elastic net



Source: Authors' construct from the data

The predictive performance of the three models was evaluated based on the optimal value of the penalty term (lambda). LASSO outperformed the other two techniques, as evidenced by the lowest mean square error value of 2.017042 as against 2.022117 and 2.023353 for ridge and elastic net, respectively. The coefficient of the optimal model presented in Table 1 suggests that the number of predictors retained by lasso and elastic net were almost alike except that rainfall was retained in lasso while it shrank to zero in elastic net.

Table 1: Coefficient of the optimal model for each algorithm

	RIDGE	LASSO	ENET	RANDOM F
(Intercept)	-1.06526853	-1.06427631	-1.06548993	
Aridity	-0.08086620	-0.05088463	-0.05137154	
Enhanced Vegetation	0.06322053	0.06622958	0.07856605	

Index				
ITN Coverage	0.03418025	0.03427459	0.03730621	
Livestock Cattle	-0.00438099			
Livestock Chickens	-0.01699454	-0.00851419	-0.01092425	
Livestock Goats	-0.02213552	-0.02239375	-0.02113944	
Livestock Sheep	-0.00311612			
Maximum Temperature	0.07813107	0.08358446	0.10248554	
Minimum Temperature	-0.10656113	-0.10923143	-0.13127881	
Proximity to Water	-0.02276671	-0.01701567	-0.03306583	
Rainfall	0.02102439	.	0.00376650	
Nightlights Composite	-0.05106234	-0.04694075	-0.04433307	
Slope	-0.01554965	-0.00633485	-0.01172392	

Source: Authors' construct from the data

The result obtained with the regularised algorithms was compared with ensemble Poisson random forest feature importance. Though the performance metric of random forest is far better than all the regularisation techniques (r-square 0.612097 versus 0.302387 for ridge, 0.309929 for lasso, and 0.315714 for elastic net), the number of selected predictors was exactly the same as that of lasso. The noticeable differences are the kinds of variables, which were mainly caused by the behavior of the two techniques in the presence of strongly correlated predictors which can be explored further in future research.

Discussion of Result

The discussion section critically evaluates the study's findings, with a focus on how successfully LASSO, Ridge, and Elastic Net conduct feature selection for malaria spread. This section compares the advantages and disadvantages of each approach, the relevant variables revealed, and the implications of these findings for improving malaria prediction models and effecting public health actions using mean square errors and R-squared values.

Overview of Key Findings

The study compared the performance of three machine learning methods—LASSO, Ridge, and Elastic Net—in detecting critical indicators of malaria transmission. With the lowest mean square error (MSE) of 2.017042, LASSO outperformed the other regularisation methods, including Elastic Net (MSE = 2.023353) and Ridge (MSE = 2.0222117). The 10 important variables identified by LASSO as having the greatest influence on malaria transmission were aridity, enhanced vegetation index, ITN coverage, livestock chickens and goats, maximum and minimum temperatures, proximity to water, nightlight composite, and slope. These variables demonstrate the complexities of malaria transmission patterns by identifying a combination of environmental, socioeconomic, and infrastructure factors. These critical components can be identified while minimising multicollinearity across predictors, due to LASSO's capacity to select variables by reducing coefficients to zero.

Although the reduced MSE indicated that LASSO surpassed Ridge and Elastic Net in prediction accuracy, the Poisson-based Random Forest technique generated even more promising results. With a substantially lower MSE of 0.006167, the Random Forest model appeared to outperform in terms of prediction accuracy. It also showed a higher R-squared value, indicating that it has greater explanatory power and can capture complicated, nonlinear interactions between factors. Even though the Random Forest model performs somewhat better, LASSO is a significant tool for malaria research because it allows for interpretability by picking a smaller sample of important predictors, which is critical for a clear understanding of the major drivers. This study highlights the combined benefits of regularisation and ensemble methods in disease modelling and demonstrates how machine learning methodologies may be utilised to build malaria prediction models.

Comparison with Previous Literature

The findings of this study, which reveal that Poisson-based Random Forest outperforms Ridge and Elastic Net in terms of prediction accuracy and that LASSO outperforms Ridge and Elastic Net in predicting malaria spread, should be compared to previous work in the larger machine learning literature. This section critically evaluates previous works that used these regularisation techniques in the context of disease modelling, specifically malaria transmission, to draw relevant scholarly conclusions regarding their efficacy and limitations.
LASSO in Disease Prediction: A Common Choice for Variable Selection

LASSO is frequently used in epidemiological studies and illness prediction due to its effectiveness in variable selection, especially when multicollinearity is present. Only the most significant predictors for the model are produced through the LASSO regularisation technique, which penalises the absolute value of the coefficients to prevent overfitting(Greenwood et al., 2020).This approach has been utilised successfully in a variety of disease modelling settings, including malaria prediction.Yamba et al., (2023) employed LASSO to anticipate malaria transmission in sub-Saharan Africa, with temperature and rainfall serving as significant environmental and climatic parameters. Researchers were able to make clear conclusions about how climate factors affect malaria outbreaks because of LASSO's capacity to manage multicollinearity and its comparatively high interpretability(Ryan et al., 2020). Similarly, Amadi & Erandi, (2024) used LASSO in conjunction with climate data to forecast the incidence of seasonal malaria in Senegal and discovered that it was a good tool for identifying key variables.

While this study corroborates previous findings about LASSO's efficacy in predicting malaria spread, it also makes an important observation: even though LASSO outperformed Ridge and Elastic Net in terms of mean square error (MSE) and successfully identified significant predictors, its overall performance was not significantly superior. Zhao et al. (2016) addressed a subtle but important point when they stated that, while LASSO gives unambiguous variable selection, it may not always produce noticeably better predictions than other approaches, particularly when the underlying relationships are complex and non-linear(Huang et al., 2024). This study's finding that Random Forest outperformed LASSO in terms of prediction accuracy, as evidenced by a higher R-squared and significantly lower MSE, is consistent with Khan et al., (2024) and suggests that the effectiveness of ensemble approaches may outweigh LASSO's emphasis on variable selection in some cases.

Ridge Regression: Performance and Usefulness in Disease Modelling

Ridge regression is a well-known regularisation technique that, unlike LASSO, penalises the square of the coefficients rather than the absolute value. This keeps all predictors, albeit in lesser magnitudes, by forcing the coefficients to shrink towards zero without totally zeroing them and this is especially useful when there is multicollinearity and all predictors are deemed to have some predictive value(Schreiber-Gregory, 2018). In a study by Yadav & Sharma, (2022) in India using Ridge regression in the field of malaria modelling to investigate the association between environmental parameters and malaria

incidence. Although ridge regression had the disadvantage of having more variables than needed, which could result in less interpretable results, it was found to operate effectively when elements were closely related(Davis et al., 2019). Despite doing somewhat worse than LASSO in this study, Ridge was still able to detect crucial elements that contribute to malaria transmission, such as socioeconomic and environmental factors.

This confirms the current study's outcome that Ridge is still a reasonable alternative for ensuring that all variables are considered rather than just one, even though it performs worse than LASSO in terms of MSE. The study's findingsthat Ridge and Elastic Net performed similarly to LASSO, imply that Ridge could be a valuable tool in other instances where new predictors are expected to gradually contribute to the prediction task(Ahmed et al., 2022). However, LASSO is preferred for malaria transmission, which requires a smaller collection of highly significant predictors; The findings reported here call into question the concept that Ridge should always function on par with LASSO, highlighting the importance of thoroughly investigating disease dynamics and accessible information in each situation(Aheto et al., 2021).

Elastic Net: Combining Strengths of LASSO and Ridge

Elastic Net attempts to strike a balance between variable selection and coefficient shrinkage by combining Ridge and LASSO's L1 and L2 penalties. It has been shown that when predictors are heavily correlated, it performs well since it prefers to select groups of comparable forecasters rather than individual predictors. Because of its capacity to address multicollinearity while maintaining interpretability, elastic net is an attractive replacement for representing complicated diseases such as malaria(Greenwood et al., 2020).

Bailey & Prist, (2024) used Elastic Net to forecast the spread of vector-borne illnesses in tropical regions, including malaria. They found that Elastic Net outperformed LASSO and Ridge when environmental and socioeconomic factors were significantly correlated. This contrasts with the findings of this study, which found that while Elastic Net could handle correlated predictors, LASSO performed better. This mismatch is caused by the specificity of the malaria transmission statistics employed in the study. A clear set of ten predictors that did not necessitate the type of inter-correlation handling that Elastic Net provides influenced malaria transmission in the current study, making LASSO a more basic and efficient alternative(Aheto et al., 2021).

Furthermore, the results show that, while Elastic Net works well in some cases, it may not always outperform LASSO or Ridge in instances when the dataset is constrained or the interactions between predictors are straightforward. This observation is consistent with the findings of Hastie et al. (2015), who stated that while Elastic Net's benefits become more obvious in high-dimensional datasets, simpler approaches such as LASSO or Ridge can be adequate in smaller, easier-to-manage datasets(Mueni, 2022).

Critique and Synthesis

The study's core finding, that LASSO is the best regularisation approach for predicting malaria spread, adds to the expanding body of evidence supporting LASSO's efficacy in epidemiological modelling(Chen et al., 2018). Nonetheless, Random Forest's performance in this work contradicts the notion that regularisation techniques are usually superior for prediction tasks involving complex disorders(Farhadi et al., 2022). Future malaria transmission research should focus on ensemble approaches, as Random Forest's improved performance demonstrates its capacity to capture complicated, non-linear interactions that regularisation techniques may overlook(Lydia & Chandrasekar, 2022).

However, the study's findings highlight fundamental challenges about how to balance interpretability with model complexity. LASSO provides a simple and interpretable model with a manageable number of predictors; however, Random Forest's better prediction accuracy compromises this interpretability, which is critical for public health decision-making. A fundamental problem in machine learning applications to malaria modelling and, more broadly, infectious disease epidemiology is the tension between model interpretability and accuracy(Chen et al., 2018).

The findings of this work imply that LASSO is an excellent strategy for discovering key variables in malaria transmission models, particularly when multicollinearity is present. Nonetheless, Random Forest's higher performance highlights the importance of researching nonlinear approaches in future research. More precise and understandable malaria prediction models may be achieved by combining multiple models or regularisation techniques with ensemble methods(Carneiro et al., 2022). Future study should look into how these approaches might be coupled to balance model interpretability with prediction accuracy, which will lead to more successful malaria control tactics.

Conclusion

This study examined the performance of machine learning regularisation techniques—LASSO, Ridge, and Elastic Net—in forecasting malaria spread, with a focus on finding significant environmental and socioeconomic aspects. LASSO was shown to be the most successful of the three regularisation strategies in terms of mean square error (MSE). It chose eleven significant characteristics that increase the risk of malaria, including temperature changes, aridity, vegetation indices, and proximity to water. Despite LASSO's performance, a Poisson-based Random Forest model outperformed it in terms of prediction accuracy, with a substantially lower MSE and more explanatory power. This demonstrates the utility of ensemble approaches for difficult, nonlinear sickness prediction problems.

The study's main finding is that, while regularisation techniques such as LASSO are effective at identifying important predictors in disease modelling, ensemble techniques such as Random Forest have a higher predictive capacity for detecting complex, non-linear correlations in malaria transmission data. These findings underscore the importance of striking a balance between model interpretability and accuracy, especially in public health applications. Future research should include hybrid modelling approaches that combine the predictive capability of ensemble methods with the interpretative clarity of regularisation techniques to improve the accuracy of malaria forecasting and other epidemiological models.

References

Adamua, Y. A., & Singh, J. (2021). Malaria Prediction Model Using Machine Learning Algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 7488–7496. https://doi.org/10.17762/turcomat.v12i10.5655

Agrawal, S. (2023). Feature Selection Using Lasso Regression. https://medium.com/@agrawalsam1997/feature-selection-using-lasso-regression-10f49c973f08

Aheto, J. M. K., Duah, H. O., Agbadi, P., & Nakua, E. K. (2021a). A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare? Preventive Medicine Reports, 23(February), 101475. https://doi.org/10.1016/j.pmedr.2021.101475

Aheto, J. M. K., Duah, H. O., Agbadi, P., & Nakua, E. K. (2021b). A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare? Preventive Medicine Reports, 23. https://doi.org/10.1016/j.pmedr.2021.101475

Ahmed, M. A. A., Sharma, E., Janifer Jabin Jui, S., Deo, R. C., Nguyen-Huy, T., & Ali, M. (2022). Kernel Ridge Regression Hybrid Method for Wheat Yield Prediction with Satellite-Derived Predictors. Remote Sensing, 14(5), 1136. https://doi.org/10.3390/rs14051136

Amadi, M., & Erandi, K. K. W. H. (2024). Assessing the relationship between malaria incidence levels and meteorological factors using cluster-integrated regression. BMC Infectious Diseases, 24(1), 1–17. https://doi.org/10.1186/s12879-024-09570-z

Awasthi, K. R., Jancey, J., Clements, A. C. A., Rai, R., & Leavy, J. E. (2024). Community engagement approaches for malaria prevention, control and elimination: a scoping review. BMJ Open, 14(2), 1–17. https://doi.org/10.1136/bmjopen-2023-081982

Bailey, A., & Prist, P. R. (2024). Landscape and Socioeconomic Factors Determine Malaria Incidence in Tropical Forest Countries. International Journal of Environmental Research and Public Health, 21(5), 576. https://doi.org/10.3390/ijerph21050576

Carneiro, T. C., Rocha, P. A. C., Carvalho, P. C. M., & Fernández-Ramírez, L. M. (2022). Ridge regression ensemble of machine learning

models applied to solar and wind forecasting in Brazil and Spain. Applied Energy, 314. https://doi.org/10.1016/j.apenergy.2022.118936
CDC. (2024). Malaria's Impact Worldwide. https://www.cdc.gov/malaria/php/impact/index.html

Chan, J. Y. Le, Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. In Mathematics (Vol. 10, Issue 8, p. 1283). Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/math10081283

Chen, Y., Chu, C. W., Chen, M. I. C., & Cook, A. R. (2018). The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. Journal of Biomedical Informatics, 81, 16–30. https://doi.org/10.1016/j.jbi.2018.02.014

Cule, E., & De Iorio, M. (2013). Ridge regression in prediction problems: Automatic choice of the ridge parameter. Genetic Epidemiology, 37(7), 704–714. https://doi.org/10.1002/gepi.21750

Davis, J. K., Gebrehiwot, T., Worku, M., Awoke, W., Mihretie, A., Nekorchuk, D., & Wimberly, M. C. (2019). A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model. Environmental Modelling and Software, 119, 275–284. https://doi.org/10.1016/j.envsoft.2019.06.010

De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. Journal of Complexity, 25(2), 201–230. https://doi.org/10.1016/j.jco.2009.01.002

Farhadi, Z., Bevrani, H., & Feizi-Derakhshi, M. R. (2022). Combining Regularization and Dropout Techniques for Deep Convolutional Neural Network. IEEE Global Energy Conference, GEC 2022, 335–339. https://doi.org/10.1109/GEC55014.2022.9986657

Filzmoser, P., & Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. Wiley Interdisciplinary Reviews: Computational Statistics, 13(4), 1–18. https://doi.org/10.1002/wics.1524

Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of Support Vector Machine Algorithm in Big Data Background. Mathematical Problems in Engineering, 2021. https://doi.org/10.1155/2021/5594899

Gimba, B., & Bala, S. I. (2017). Modeling the Impact of Bed-Net Use and Treatment on Malaria Transmission Dynamics. International

Scholarly Research Notices, 2017, 1–16.
https://doi.org/10.1155/2017/6182492

Greenwood, C. J., Youssef, G. J., Letcher, P., Macdonald, J. A., Hagg, L. J., Sanson, A., McIntosh, J., Hutchinson, D. M., Toumbourou, J. W., Fuller-Tyszkiewicz, M., & Olsson, C. A. (2020). A comparison of penalised regression methods for informing the selection of predictive markers. PLoS ONE, 15(11 November), 97–108.
https://doi.org/10.1371/journal.pone.0242730

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., & Hao, Y. (2015). Improved variable selection algorithm using a LASSO-Type penalty, with an application to assessing hepatitis b infection relevant factors in community residents. PLoS ONE, 10(7), 1–23.
https://doi.org/10.1371/journal.pone.0134151

Huang, Y., Tibbe, T., Tang, A., & Montoya, A. (2024). Lasso and Group Lasso with Categorical Predictors: Impact of Coding Strategy on Variable Selection and Prediction. Journal of Behavioral Data Science, 3(2), 15–42. https://doi.org/10.35566/jbds/v3n2/montoya

Khan, M. A., Azim, A., Liscano, R., Smith, K., Chang, Y. K., Seferi, G., & Tauseef, Q. (2024). On the Effectiveness of Feature selection Techniques in the Context of ML-based Regression Test Prioritization. IEEE Access, 12(August), 131556–131575.
https://doi.org/10.1109/ACCESS.2024.3459656

Khan, O., Ajadi, J. O., & Hossain, M. P. (2024). Predicting malaria outbreak in The Gambia using machine learning techniques. PLoS ONE, 19(5), e0299386. https://doi.org/10.1371/journal.pone.0299386

Kino, S., Hsu, Y. T., Shiba, K., Chien, Y. S., Mita, C., Kawachi, I., & Daoud, A. (2021). A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. SSM - Population Health, 15, 100836.
https://doi.org/10.1016/J.SSMPH.2021.100836

Kyriazos, T., & Poga, M. (2023). Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. Open Journal of Statistics, 13(03), 404–424. https://doi.org/10.4236/ojs.2023.133020

Li, Y., & Abdallah, S. (2022). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316.

https://www.sciencedirect.com/science/article/pii/S0925231220311693

Li, Z., Ye, C. Y., Zhao, T. Y., Zhao, T. Y., & Yang, L. (2020). Model of genetic and environmental factors associated with type 2 diabetes mellitus in a Chinese Han population. BMC Public Health, 20(1), 1–12. https://doi.org/10.1186/s12889-020-09130-5

Lucas, T. C. D., Nandi, A. K., Keddie, S. H., Chestnutt, E. G., Howes, R. E., Rumisha, S. F., Arambepola, R., Bertozzi-Villa, A., Python, A., Symons, T. L., Millar, J. J., Amratia, P., Hancock, P., Battle, K. E., Cameron, E., Gething, P. W., & Weiss, D. J. (2022). Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence. Spatial and Spatio-Temporal Epidemiology, 41, None. https://doi.org/10.1016/j.sste.2020.100357

Lydia, A. A., & Chandrasekar, S. (2022). A Comparative Study on Regularization Techniques in Convolutional Neural Networks. In International Journal of Research in Engineering and Science (IJRES) ISSN (Vol. 10). https://www.researchgate.net/publication/362279351_A_Comparative_Study_on_Regularization_Techniques_in_Convolutional_Neural_Networks

Mueni, M. (2022). A Systematic comparison of performance of Ridge , Lasso , Elastic net and Relaxed Elastic net when fitting high dimensional data for sales prediction . A Systematic Comparison of Performance of Ridge , Lasso , Dimensional Data for Sales Prediction.

Mujahid, M., Rustam, F., Shafique, R., Montero, E. C., Alvarado, E. S., de la Torre Diez, I., & Ashraf, I. (2024). Efficient deep learning-based approach for malaria detection using red blood cell smears. Scientific Reports, 14(1), 1–16. https://doi.org/10.1038/s41598-024-63831-0

Muthukrishnan, R., & Rohini, R. (2017). LASSO: A feature selection technique in predictive modeling for machine learning. 2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016, 18–20. https://doi.org/10.1109/ICACA.2016.7887916

Nilashi, M., Keng Boon, O., Tan, G., Lin, B., & Abumalloh, R. (2023). Critical Data Challenges in Measuring the Performance of Sustainable Development Goals: Solutions and the Role of Big-Data Analytics. Harvard Data Science Review, 5(3). https://doi.org/10.1162/99608f92.545db2cf

Nugroho, H. (2023). A Review: Data Quality Problem in Predictive Analytics. IJAIT (International Journal of Applied Information Technology), 7(02), 79. https://doi.org/10.25124/ijait.v7i02.5980

Obadina, O. G., Adedotun, A. F., & Odusanya, O. A. (2021). Ridge Estimation's Effectiveness for Multiple Linear Regression with Multicollinearity: An Investigation Using Monte-Carlo Simulations. Journal of the Nigerian Society of Physical Sciences, 3(4), 278–281. https://doi.org/10.46481/jnsps.2021.304

Obasohan, P. E., Walters, S. J., Jacques, R., & Khatab, K. (2021). A scoping review of selected studies on predictor variables associated with the malaria status among children under five years in sub-Saharan Africa. International Journal of Environmental Research and Public Health, 18(4), 1–21. https://doi.org/10.3390/ijerph18042119

Ochieng, F. O. (2024). SEIRS model for malaria transmission dynamics incorporating seasonality and awareness campaign. Infectious Disease Modelling, 9(1), 84–102. https://doi.org/10.1016/j.idm.2023.11.010

Okunlola, O. A., Oyeyemi, O. T., & Lukman, A. F. (2021). Modeling the relationship between malaria prevalence and insecticide-treated bed net coverage in Nigeria using a Bayesian spatial generalized linear mixed model with a Leroux prior. Epidemiology and Health, 43, e2021041. https://doi.org/10.4178/EPIH.E2021041

Olushola, A., Mart, J., & Alao, V. (2023). Predictive Modelling For Disease Outbreak (Vol. 1). https://www.researchgate.net/publication/377777597_PREDICTIVE_MODELLING_FOR_DISEASE_OUTBREAK_PREDICTION

Orimadegun, A. E., & Ilesanmi, K. S. (2015). Mothers' understanding of childhood malaria and practices in rural communities of Ise-Orun, Nigeria: implications for malaria control. Journal of Family Medicine and Primary Care, 4(2), 226. https://doi.org/10.4103/2249-4863.154655

Raiaan, M. A. K., Sakib, S., Fahad, N. M., Mamun, A. Al, Rahman, M. A., Shatabda, S., & Mukta, M. S. H. (2024). A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks. In Decision Analytics Journal (Vol. 11, p. 100470). Elsevier. https://doi.org/10.1016/j.dajour.2024.100470

Rauschenberger, A., Glaab, E., & Van De Wiel, M. A. (2021). Predictive and interpretable models via the stacked elastic net. Bioinformatics, 37(14), 2012–2016. https://doi.org/10.1093/bioinformatics/btaa535

Ryan, S. J., Lippi, C. A., & Zermoglio, F. (2020). Shifting transmission risk for malaria in Africa with climate change: A framework for planning and intervention. Malaria Journal, 19(1), 1–14. https://doi.org/10.1186/s12936-020-03224-6

Schreiber-Gregory, D. N. (2018). Ridge Regression and multicollinearity: An in-depth review. Model Assisted Statistics and Applications, 13(4), 359–365. https://doi.org/10.3233/MAS-180446

Shekarau, E., Uzoanya, M., Ogbulafor, N., Ntadom, G., Ijezie, S. N., Uzoanya, M. I., Seye, B., Fashanu, C., Eze, N., Nwidae, L., Mokuolu, O., Nwokenna, U., Nglass, I., Ishola-Gbenla, O., Okouzi, M., Fagbola, M., Oresanya, O., Getachew, D., Chukwumerije, J., … Rietveld, H. (2024). Severe malaria intervention status in Nigeria: workshop meeting report. Malaria Journal, 23(1). https://doi.org/10.1186/s12936-024-05001-1

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8(2), 39–42. https://doi.org/10.12691/ajams-8-2-1

Shretta, R., Liu, J., Cotter, C., Cohen, J., Dolenz, C., Makomva, K., Newby, G., Ménard, D., Phillips, A., Tatarsky, A., Gosling, R., & Feachem, R. (2017). Malaria Elimination and Eradication. In Disease Control Priorities, Third Edition (Volume 6): Major Infectious Diseases (pp. 315–346). The International Bank for Reconstruction and Development / The World Bank. https://doi.org/10.1596/978-1-4648-0524-0_ch12

Stanley, C. C., Kazembe, L. N., Mukaka, M., Otwombe, K. N., Buchwald, A. G., Hudgens, M. G., Mathanga, D. P., Laufer, M. K., & Chirwa, T. F. (2019). Systematic review of analytical methods applied to longitudinal studies of malaria. Malaria Journal, 18(1), 254. https://doi.org/10.1186/s12936-019-2885-9

Usman, M., Doguwa, S. I. S., & Alhaji, B. B. (2022). Comparing the Prediction Accuracy of Ridge, Lasso and Elastic Net Regression Models with Linear Regression Using Breast Cancer Data. Bayero Journal of Pure and Applied Sciences, 14(2), 134–149. https://doi.org/10.4314/bajopas.v14i2.16

Vatcheva, P. K., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology: Open Access, 06(02), 227. https://doi.org/10.4172/2161-1165.1000227

WHO. (2022). Report on malaria in Nigeria 2022 | WHO | Regional Office for Africa. https://www.afro.who.int/countries/nigeria/publication/report-malaria-nigeria-2022-0

Wiemken, T. L., & Kelley, R. R. (2019). Machine learning in epidemiology and health outcomes research. Annual Review of Public Health, 41, 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437

Yadav, C. P., & Sharma, A. (2022). National Institute of Malaria Research-Malaria Dashboard (NIMR-MDB): A digital platform for analysis and visualization of epidemiological data. The Lancet Regional Health - Southeast Asia, 5, 100030. https://doi.org/10.1016/j.lansea.2022.100030

Yamba, E. I., Fink, A. H., Badu, K., Asare, E. O., Tompkins, A. M., & Amekudzi, L. K. (2023). Climate Drivers of Malaria Transmission Seasonality and Their Relative Importance in Sub-Saharan Africa. GeoHealth, 7(2). https://doi.org/10.1029/2022GH000698

Yang, X., & Wen, W. (2018). Ridge and Lasso Regression Models for Cross-Version Defect Prediction. IEEE Transactions on Reliability, 67(3), 885–896. https://doi.org/10.1109/TR.2018.2847353

Yazdi, M., Golilarz, N. A., Nedjati, A., & Adesina, K. A. (2021). An improved lasso regression model for evaluating the efficiency of intervention actions in a system reliability analysis. Neural Computing and Applications, 33(13), 7913–7928. https://doi.org/10.1007/s00521-020-05537-8

Ying, X. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series, 1168(2). https://doi.org/10.1088/1742-6596/1168/2/022022

Yu, T., & Zhu, H. (2020). Hyper-Parameter Optimization: A Review of Algorithms and Applications. http://arxiv.org/abs/2003.05689