Comparing the Effectiveness of Support Vector Regression, Random Forest and Lasso Regression in Forecasting Dengue Cases

Deniz Sagmanli
School of Computer Science
University of Nottingham
Nottingham, UK
psxds8@nottingham.ac.uk

Muhammad Hassan Rizvi
School of Computer Science
University of Nottingham
Nottingham, UK
psxmr4@nottingham.ac.uk

Sundeep Veluchamy
School of Computer Science
University of Nottingham
Nottingham, UK
psxsv2@nottingham.ac.uk

***Abstract*— Dengue is an endemic tropical disease spread by Aedes aegypti mosquitoes. This study attempted to model and understand dengue spread in two cities San Juan, Puerto Rico and Iquitos, Peru. It attempted to understand and predict the weekly dengue cases and dengue outbreaks in the city by comparing the Random Forest Regression, Support Vector Regression and Lasso Regression. The study used Bayesian hyperparameter optimization and tried to understand the effect of changing the scoring metric for the hyperparameter optimization algorithm from mean absolute error to coefficient of determination and its effect on the predictive ability of the produced models. Finally, it also used a feature importance analysis to extract the most important features and attempted to understand the effect of increasing the number of features (from 6 features to 10 features) and its effect on the predictive ability of the models. The results were analysed using various average performance metrics (Mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination (R^2)) obtained through the nested cross-validation methodology and confidence intervals were provided for these metrics. For the city of San Juan the Support Vector Regression produced more promising performance metrics but the random forests model was able to predict dengue outbreaks with a much greater degree of accuracy. In general, for both cities changing the scoring metric to R^2 the hyperparameter optimization diminished the performance metrics of the models that had performed well under the MAE scoring metric and did not in general affect or improve any of the remaining models. A similar result was obtained when the number of features used to create the machine learning model was increased from 6 to 10.**

Keywords—Dengue, Mean Absolute Error, Root Mean Squared Error, R^2 Score, Support Vector Machine, Random Forest, Lasso Regression.

I. Introduction

Dengue fever is a mosquito-borne disease prevalent in tropical and subtropical regions, posing a significant global health challenge. Its transmission dynamics are intricately linked to climate variables, particularly precipitation and temperature [1]. The union between climate change and dengue accentuates the urgency of understanding its epidemiology and devising predictive models to mitigate its impact.

The Dataset under study encompasses the epidemiological data of dengue fever cases across two regions: Iquitos and San Juan. Weekly records of fever cases alongside climate variables are recorded in detail for multiple years enabling in-depth analysis of the disease's dynamic nature [2].

This research aims to address three critical questions: Firstly, it delves into the efficacy of three machine-learning regression models– Support Vector Machines (SVM), Lasso Regression, and Random Forest– in predicting the weekly dengue cases. The aim is to identify which model is the most suitable for this task of prediction in each city by evaluating each model's forecasting accuracy and robustness. The second inquiry revolves around the impact of altering the scoring metric of Bayesian Hyperparameter Optimization from Mean absolute error to the coefficient of determination and its impact on model performance. This exploration seeks to anticipate the optimal configuration for the selected regression models, a more thorough investigation is required in the future in this area. The last investigation involves exploring the influence of feature dimensionality on the model's resilience. Employing a stepwise feature selection approach, additional features shall be incorporated incrementally from the dataset which will help assess the trade-off between model complexity and generalisation performance. By analysing this relationship, it will be easy to identify the optimal feature set that facilitates the development of accurate and resilient models for predicting dengue fever incidence in the specific cities involved in this study.

II. Literature Review

Various forecasting techniques have been used to model Dengue fever for different purposes. Machine learning models have been used to predict the likelihood and location of Dengue cases [3], they have also been used in conjunction with other methods such as agent-based simulations to understand and predict the effectiveness of various mitigation strategies on the spread of mosquito-borne diseases similar to Dengue [4]. Time series forecasting techniques such as seasonal ARIMA and Neural networks have also been used to predict Dengue incidence [5].

The literature suggests that there does not seem to be a particular type of model that generally models the weekly dengue cases uniquely well. Reference [6] suggests that Linear Models can potentially outperform Support Vector regression in terms of Mean Absolute Error in the context of their research which modelled Dengue spread in the cities in Bangladesh. Their research took into account trends in hospitalised patient data and socio-economic data as well.

	Feature Name	Coefficient	Absolute Coefficient
0	sin_weekofyear	-10.952785	10.952785
1	reanalysis_specific_humidity_g_per_kg	6.060097	6.060097
2	year	-2.302270	2.302270
3	ndvi_ne	-1.694585	1.694585
4	ndwi_nw	-0.664662	0.664662
5	reanalysis_precip_amt_kg_per_m2	0.618205	0.618205
6	ndvi_se	-0.137520	0.137520
7	reanalysis_avg_temp_k	0.000000	0.000000
8	reanalysis_dew_point_temp_k	0.000000	0.000000
9	reanalysis_max_air_temp_k	0.000000	0.000000
10	reanalysis_min_air_temp_k	0.000000	0.000000
11	reanalysis _air_temp_k	0.000000	0.000000
12	ndvi_sw	0.000000	0.000000

TABLE I
THE RESULTS OF THE LASSO FEATURE ANALYSIS

However, another study that examined different machine learning models and feature selection strategies suggested that, in general, the random forests models trained on monthly dengue cases outperformed (in mean absolute error) support vector regression, gradient boosting regression, feed-forward neural networks and a naive Bayes Model [7]. They also concluded that different models worked best in different cities. This corroborates the findings [8]. They used multiple techniques to try and predict weekly dengue cases who concluded that of the 9 cities they studied (in the state of Maharashtra, India) Random Forest performed the best in 5 cities and support vector regression in 2 cities [8].

A more pertinent piece of research was regarding a very similar piece of research on the cities of San Juan and Iquitos (and Singapore) where they attempted to create a model to forecast weekly Dengue cases [9]. However, in the paper they compared random forest regression with random forest Univariate Flagging algorithm and Time Series Models such as ARIMA and seasonal ARIMA. Poisson Regression, Logistic Regression were also compared together with the LASSO algorithm. The results seemed to show that RF had lesser mean absolute error than Poisson, Logistic and ARIMA models [9]. It was conjectured in the paper that RF's effectiveness in modelling Dengue could be due to its ability to model non-linear relationships effectively [9]. However they had not compared the effectiveness of Support Vector Regression in predicting weekly dengue cases in the two Latin American cities, which is an area of exploration we explore in the paper.

III. Methodology

A short summary of the methodology can be found in Figure 1.

A. Data Cleaning and Imputation

The dataset collected from [2] had missing values, particularly in the vegetation columns, these missing values were addressed by employing the moving average method of imputation, which essentially calculates the rolling averages considering the temporal dynamics of the data. The dataset was split into two subsets based on the two cities: San Juan and Iquitos, the missing values for the two data subsets were imputed individually ensuring these imputations reflect each city's unique epidemiological profiles. Additionally, to account for the seasonality of the dataset a sine transformation was applied to the 'week of year' column/feature using sine and cosine functions, which resulted in enhancing the models' ability to capture weekly fluctuations in dengue incidences much more efficiently. Feature extraction and selection were conducted using Lasso Linear Regression, which promotes sparsity in the feature space and identifies the most influential predictors. Lasso regression's L1 regularisation reduces the coefficient of the less informative variables towards zero, and vice versa [10], improving the model's interpretability and efficiency while reducing overfitting [11]. This study employed Linear lasso regression for feature extraction and investigated the impact of increasing feature numbers on model robustness. The results of the feature analysis are presented in Table I.

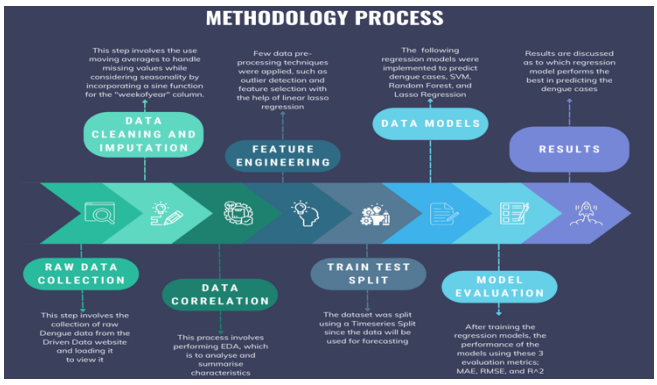

Fig. 1. A summarised flowchart of the methodology process.

This study utilised a factorial design with 6 and 10 features. It was found that only 7 features had non-zero coefficients as per the Lasso Feature analysis. To enrich the feature set, domain knowledge aided in the inclusion of three more features such as "reanalysis_relative_humidity_percent, reanalysis_sat_percip_amt_mm, and station_avg_temp_c". Despite these features having zero coefficients initially as per the Lasso Feature analysis, these variables are crucial in dengue transmission dynamics, which enhance the model's depth and align with the study's goal of evaluating feature selection's impact on model efficiency.

B. Feature Extraction

Having quality data is of utmost importance when developing predictive models so that accurate results can be computed efficiently. To improve the quality of the data, feature extraction/selection was conducted. This process of selecting the most relevant features in the dataset helps in training the predictive models much faster and reduces the overfitting of the model while also improving the accuracy of the predictive models. Linear Lasso regression is employed for feature selection, aiming to identify the most influential predictors. Initially we attempted to use Random Forest feature importance analysis as Random forest regression can potentially capture many non-linear dependencies more accurately. But we decided to use lasso regression, lasso regression uses the L1 regularisation to promote sparsity in the feature space. This helps in feature selection by reducing the coefficients of the less informative variables towards zero, simultaneously, focusing on the subset of relevant features which remain non-zero after the L1 penalty which are considered to have a significant impact on the target variable [10].

Lasso regression improves the predictive model's interpretability and reduces the overfitting of the model, ultimately leading to a more efficient and interpretable model [11]. In addition to the feature extraction through Lasso regression, an examination of the effect of increasing the number of features is conducted in this study to test the hypothesis that a few features can suffice in building a robust model. Utilizing factorial design with 6 and 10 features.

It is to be noted that only 7 non-zero coefficient features were extracted using Lasso feature analysis, to incorporate additional features, domain knowledge was leveraged, leading to the inclusion of further three important variables: "reanalysis_relative_humidity_percent, reanalysis_sat_percip_amt_mm, and station_avg_temp_c" despite having coefficients of 0 in the current analysis, these features play crucial roles in the transmission of Dengue. Relative humidity and precipitation amount significantly influence mosquito breeding habitats, while temperature affects mosquito survival and virus replication rates. Incorporating these domain-relevant features enhances the predictive model's comprehensiveness and aligns with the study's objective of assessing the impact of feature selection on model robustness and efficiency.

C. Predictive Modelling

Three predictive modelling techniques are utilised in this study which are, support vector Regression (SVR), Random Forest, and Lasso Regression. Each of these models offers unique methodologies for forecasting Dengue incidences. While SVR seeks to pinpoint the optimal hyperplane given a certain allowed epsilon margin from points in the dataset, Lasso Regression performs linear regression with L1 regularisation, prioritising only those features that are the most influential [10] and lastly, Random Forest leverages ensemble learning techniques to accumulate predictions from multiple decision trees, resulting in an enhanced model with high accuracy and robustness. A comprehensive description of how each model works is described below.

1) Support Vector Regression (SVR): In Support Vector Regression a proposed function is mapped to a high dimensional hyperplane. There exists an "**ϵ"** epsilon tube around this function based on a hyperparameter ϵ previously selected. The epsilon tube affects the sensitivity to errors since points within the epsilon tube are not penalised by the regularisation parameter. Datapoints outside the tube are used to compute the "ϵ" epsilon loss function and the goal is to minimise the objective function shown below [12].

$$\frac{1}{2}\left|\left|w\right|\right|^2 + C\ \sum_{\mu=1}^{N} |y^u - y(x^u)|_\epsilon \qquad (1)$$

The 'C' regularisation hyperparameter controls the impact of terms inside the sigma (the slack variables). The slack variables control the impact of the distance of the data points to the function on the objective function. A large 'C' value will mean that the slack variables will have a large impact on the objective function and thus smaller values for the slack variable will need to be chosen. This could potentially lead to overfitting of the training data affecting generalizability. Smaller values of the regularisation hyperparameter would have the opposite effect of having more generalizability but lesser predictive power and underfitting [12].

Another important concept that needs to be understood is the 'Kernel' of the support vector regression. It relates to the technique used to map the points in the training data to a higher dimensional space to convert it to a linear hyperplane where support vector techniques can be applied. In this study, we only consider the linear, rbf and sigmoid kernels [13].

The γ *"gamma*" hyperparameter in particular is relevant to the 'rbf' kernel; it is inversely related to the spread of the distances between the features on the features space. If gamma is low that means the kernel effect on the region around each support vector is larger therefore improving generalizability but can underfit the data. The vice versa might occur for large gamma values [12]. Finally, there is the coef0 hyperparameter which is relevant to the sigmoid kernel and affects the midpoint of the sigmoid function that is used to map the data points to the higher dimensional space [13].

2) Random Forest: An effective ensemble learning technique called Random Forest Regression builds many decision trees during training and outputs the average forecast of all the individual trees for regression tasks. When it comes to weekly Dengue case prediction, this ensemble technique has a number of benefits. The capacity of Random Forest Regression to handle huge datasets with high dimensionality is one of its main advantages; this makes it especially useful in situations where there are a lot of features or predictors. This is accomplished by randomly choosing a subset of characteristics from the decision trees at each split, which

essentially decorrelates the trees and lowers the chance of overfitting. Random Forest Regression naturally reduces the chance of bias and variance by combining predictions from several trees, producing robust and dependable predictions. Furthermore, because Random Forest Regression is ensemble in nature, it is less susceptible to noise and outliers in the data. Although a single decision tree may be prone to overfitting, the forest's average of several trees smoothes out noise and anomalies in the data, producing predictions that are more reliable and accurate.

To maximise efficiency and avoid overfitting, hyperparameters like the number of trees in the forest (n_estimators) and the maximum depth of the trees (max_depth) can be adjusted. Furthermore, by adjusting the tree development process, parameters like min_samples_split and min_samples_leaf improve the model's ability to identify underlying patterns in the data without becoming excessively complex. The built-in feature importance ranking of Random Forest Regression gives researchers the ability to determine which predictors in the dataset have the greatest influence. Researchers can learn a great deal about the factors influencing Dengue incidence by evaluating the relative importance of different traits. This knowledge can help with both prediction and interpretation.

3) Least Absolute Shrinkage and Selection Operator (LASSO) Regression: Linear Lasso regression serves a dual purpose in this study. Firstly it is used for feature extraction, and secondly, it is used as a prediction model for dengue incidences. Applying standard regression to a large pool of variables can lead to two major problems, optimism bias and overfitting.

Here, overfitting refers to the model being too complex and struggling to perform well on unseen data, optimism bias refers to the tendency of the model to overestimate its performance on new data. Both of these limitations are handled well in LASSO regression models. It is a penalised regression technique which tackles both overfitting and optimism bias simultaneously. While Lasso regression starts with the standard linear regression model, assuming a linear relationship between the independent (features) and the dependent variable (target), it further introduces an additional penalty term based on the absolute values of the coefficients. This L1 regularisation term is the sum of the absolute values of the coefficients multiplied by a new 'tuning' or 'regularisation' parameter $\boldsymbol{\lambda}$ [14]. The general equation for the Lasso Linear Regression is as follows:

$$L_1 = \lambda(|\beta_1| + |\beta_2| + \cdots + |\beta_p|) \quad (2)$$

where λ *"lambda"* is the regularisation parameter that controls the amount of regularisation applied and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients.

The objective of this technique is to compute the values for those coefficients that minimise the sum of the squared differences (RSS) between the predicted values and the actual ones, simultaneously minimising the L1 regularisation term [14]. One of the most important aspects of LASSO Regression is the choice of the regularisation parameter λ. As the value of λ increases, more coefficients are pushed towards zero and vice versa. To find the most optimal λ value K-fold cross-validation technique is implemented which helps in computing different values of λ on which the most optimum balance can be achieved between the model complexity and the prediction accuracy on unseen data.

D. Nested Cross-Validation and Hyperparameter Tuning

Hyperparameter tuning is a vital step when developing any machine learning model. Determining the most efficient hyperparameters is a tiresome and lengthy process, especially when the number of parameters which need tuning is large. Depending on the fine-tuning itself the performance of the model can be improved by 25%-90% [15]. Grid Search and Bayesian Optimization are the most common and efficient techniques employed for tuning hyperparameters. In this study, we also employ Nested Cross-Validation (CV), incorporating both time-series split and Bayesian Hyperparameter optimization.

Model	Scoring Metric Optimisation: MAE	Scoring Metric Optimisation: R^2
Random Forest	MAE: 31.2173, 95% CI = [30.7748, 31.6597] R^2: -1.9574, 95% CI = [-2.0883, -1.8265]	MAE:32.2482, 95% CI = [30.7816, 33.7148] R^2: -2.3588, 95% CI = [-2.8302, -1.8874]
Support Vector Regression	MAE: 24.4774, 95% CI = [24.4231, 24.5317] R^2: -0.2584, 95% CI = [-0.2626, -0.2541]	MAE:31.3465, 95% CI = [27.3916, 35.3013] R^2: -2.1802, 95% CI = [-3.6452, -0.7152]
Lasso Regression	MAE: 33.2098, 95% CI = [32.6801, 33.7394] R^2: -2.0571, 95% CI = [-2.3800, -1.7341]	MAE: 33.0946, 95% CI = [32.4343, 33.7549] R^2: -2.0849, 95% CI = [-2.3836, -1.7862]

TABLE II
RESULTS WITH DIFFERENT SCORING METRICS FOR ALL MODELS FOR SAN JUAN

While the time-series split ensures that the temporal dependencies within the data are appropriately accounted for since the observations are ordered chronologically. This split makes sure that the model used for training is trained on past values while being evaluated on future ones, hence, mimicking real-world forecasting scenarios. In each iteration of the nested CV, a different section of the data is selected as the test set, while the preceding data is being used for training therefore this technique adequately takes care of the problem of data leakage while providing a more realistic assessment of its performance.

The Nested CV is a cross-validation technique that attempts to overcome the problem of overfitting the training dataset. A nested CV comprises an outer and an inner loop. The outer loop is responsible for splitting the dataset into training and testing sets while the inner loop is responsible for the hyperparameter tuning. In the outer loop, the dataset is partitioned into multiple folds [16]. For each fold, the model is trained on a subset of the data and evaluated on the remaining data in the fold. Within each fold of the outer loop, the inner loop performs hyperparameter tuning using Bayesian hyperparameter optimization. Bayesian Optimization iteratively explores the hyperparameter space, selecting promising regions based on past evaluations and updating a probabilistic surrogate model of the objective function. This is done by taking into account past evaluations, Bayesian optimization efficiently narrows down the search space, ultimately identifying those hyperparameters that maximise the performance metrics. This process of iteration is continued until convergence is achieved, resulting in the selection of the most optimal hyperparameters [17]. This is done for each fold in the outer loop which leads to obtaining a set of optimal hyperparameters for each field for which we have performance metrics like MAE and Coefficient of determination.

We can average the best performance metrics on each fold to get less biased average performance metrics. It is important to note that this study uses two types of scoring metrics (different from performance metrics) to choose the best hyperparameters in the Bayesian hyperparameter optimization in each outer fold. The hyperparameter tuples are scored using the coefficient of determination (R^2 Score) and mean absolute error (MAE) separately and the best parameters are obtained for each of these scoring metrics. Future research could potentially experiment with using hybrid metrics that take into account two or more performance metrics for scoring and selecting the best parameters.

Model	Avg MAE	Avg MSE	Avg RMSE	Avg R^2
Random Forest	Mean = 31.22 95% CI = [30.77, 31.66]	Mean = 2777.63 95% CI = [2742.79, 2812.47]	Mean = 46.82, 95% CI = [46.40, 47.23]	Mean = -1.96, 95% CI = [-2.09, -1.83]
Support Vector Regression	Mean = 24.48, 95% CI = [24.42, 24.5317]	Mean = 2965.58, 95% CI = [2957.51, 2973.65]	Mean = 44.68, 95% CI = [44.60, 44.75]	Mean = -0.26, 95% CI = [-0.26, -0.25]
Lasso Regression	Mean = 33.21, 95% CI = [32.68, 33.74]	Mean = 2999.91, 95% CI = [2971.89, 3027.93]	Mean = 48.55, 95% CI = [48.03, 49.07]	Mean = -2.06, 95% CI = [-2.38, -1.73]

TABLE III
RESULTS FOR SAN JUAN WITH MAE SCORING (6 FEATURES)

Model	Avg MAE	Avg MSE	Avg RMSE	Avg R^2
Random Forest	Mean = 6.77, 95% CI = [6.73, 6.81]	Mean = 132.09, 95% CI = [131.28, 132.91]	Mean = 11.01, 95% CI = [10.97, 11.04]	Mean = -0.09, 95% CI = [-0.10, -0.08]
Support Vector Regression	Mean = 6.99, 95% CI = [6.96, 7.03]	Mean = 151.00, 95% CI = [149.49, 152.51]	Mean = 11.64, 95% CI = [11.57, 11.71]	Mean = -0.21, 95% CI = [-0.23, -0.19]
Lasso Regression	Mean = 6.53, 95% CI = [6.52, 6.53]	Mean = 142.25, 95% CI = [142.25, 142.25]	Mean = 11.34, 95% CI = [11.34, 11.34]	Mean = -0.14, 95% CI = [-0.14, -0.14]

TABLE IV
RESULTS FOR IQUITOS WITH MAE SCORING (6 FEATURES)

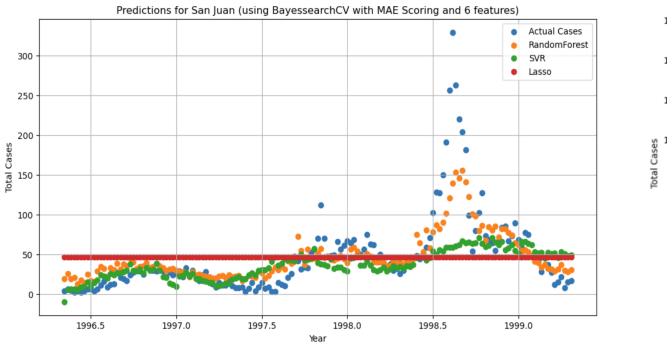

Fig. 2. A comparison of the 3 models using 6 features and MAE scoring for Bayesian hyperparameter optimisation for San Juan for years before 2000.

E. Model Evaluation

After the hyperparameter tuning, the process of nested CV is repeated 10 times with different random seeds. This generates multiple values for the average performance metrics, allowing for the computation of confidence intervals for the average performance metrics. These intervals provide insights into the variability of the average model performance, accounting for uncertainties associated with automatic Bayesian hyperparameter optimization and data variability.

F. Results

Taking into account the scoring metric for BayessearchCV and the number of features, the best model is Support Vector Regression with 6 features and MAE scoring. The results for this can be seen on Table III for San Juan and Table IV for Iquitos. An interesting finding was that the Random Forest model had a much better ability to predict outbreaks. The predictions of this result are further explored below. In general, we used the scatter plots to make comparisons in the results because of the simplicity of the performance metrics as discussed below.

Figure 2 is taken from a model in the outer loop of the nested cross-validation that was trained with data before 2005. Random forest regression seems to be able to predict outbreaks more accurately in this case but this is not reflected in the overall nested CV results in Tables III and IV, which show that the Random Forest regression performs much worse on average over the entire dataset based on the performance metrics compared to Support Vector Regression in the city of San Juan with 6 features. Figure 3 shows the R^2 Scoring metric failed to predict some of the outbreaks in this time period (more than 100 cases).

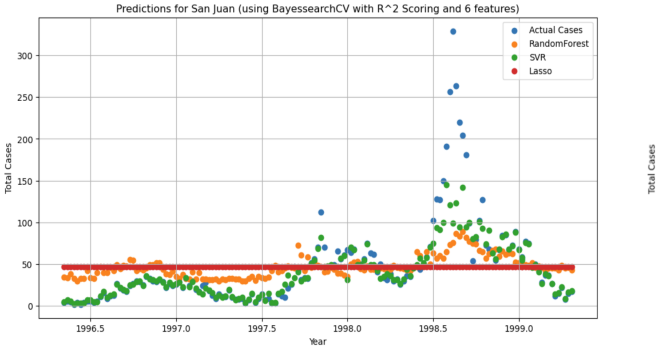

Fig. 3. A comparison of the 3 models using 6 features and R^2 scoring for Bayesian hyperparameter optimisation for San Juan for years before 2000.

This pattern is repeated in the other folds of the outer loop of the nested CV. The results were particularly poor for the city of Iquitos.

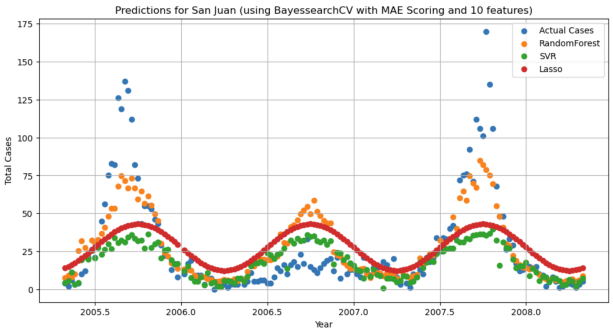

Fig. 4. A comparison of the 3 models using 10 features and MAE scoring for Bayesian hyperparameter optimisation for San Juan for years after 2005.

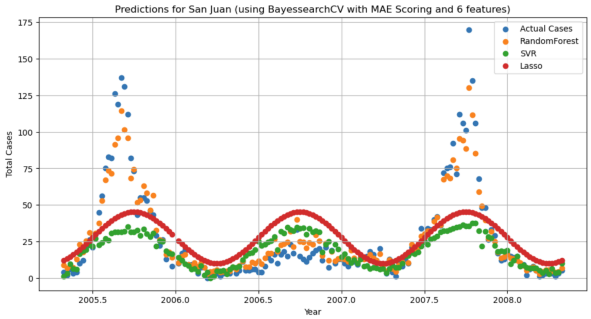

Fig. 5. A comparison of the 3 models using 6 features and MAE scoring for Bayesian hyperparameter optimisation for San Juan for years after 2005.

By looking at figure 4 and 5, increasing the number of features from 6 to 10 diminished the accuracy of the predictions for all models. These outbreaks were accurately predicted in the 6 feature models. This was also true for R^2 scoring and the city of Iquitos. This is most probably caused by the curse of dimensionality and overfitting due to having too many features.

IV. Discussion

One important idea to note when trying to interpret the results and the performance metrics is the nature of the question. A better performance in the mean absolute error and coefficient of determination (R^2) need not translate at all to a model that better predicts Dengue outbreaks. One of the results obtained by [8] is that in general, Random Forest Regression performed better in predicting daily Dengue cases and had better performance metrics than Support Vector Regression. This is not however supported by the results, as seen in Table II wherein the average mean absolute error (MAE) obtained with Support Vector Regression is nearly 20% lesser than Random Forest (obtained via nested cross-validation) in the city of San Juan. Using nested cross validation helps decrease the risk of overfitting and thus decreases bias and gives a more accurate picture of the performance metrics. Calculating average performance metrics based on Nested cross-validation (Nested CV and computing the confidence intervals by repeating this process (Nested CV) 10 times helps us obtain a much more reliable set of results that helps prevent overfitting and better understand margins of error.

There is a statistically significant difference in the predictions of the models (SVR and RF in particular) according to the tables, but the Random Forest model very clearly predicted rare event outbreaks (greater than 100 cases) in Dengue which the Support Vector Regression completely failed to do, as seen in figure 2. Support Vector Regression could therefore be potentially used to predict the changes in average cases over the long term, but Random Forest regression is likely more useful if the models are to predict the risk of outbreaks. Our results corroborate [9] and improve on their results by using different scoring metrics and models to understand Dengue outbreaks in these cities better. But another important point to note is that in the city of Iquitos, all the models seem to have performed equally badly and Random Forest too failed to predict any major outbreaks. One conjecture for the worse results for the city of Iquitos is the much smaller size of the dataset preventing good predictions over the long term.

Finally, another pattern was the consistent underperformance of Lasso Regression both in terms of performance metrics and prediction of rare event dengue outbreaks. Lasso regression may be too simplistic to model the highly non-linear dependencies between the features in dengue fever spread.

Using different scoring metrics for the Bayesian hyperparameter optimisation algorithm did influence the performance metrics of the algorithm (SVR in particular). The scoring metric is used to evaluate promising parameters in the hyperparameter space and is used by the Bayesian optimization algorithm to optimise the space further until the scores converge and an optimum set of parameters is arrived at.

The results of the nested cross-validation seem to show a worse result in the city of San Juan when the scoring metric is changed from MAE to coefficient of determination in the Support Vector model. This was not, however, true for Random Forest and Lasso Regression that had about the same performance metrics before and after in the city of San Juan. The poorer performance also extended to the prediction of dengue with the Random Forest regression which performed very well under the MAE scoring metric for the hyperparameter optimization but failed to predict some of the outbreaks, incorrectly predicted outbreaks and also gave diminished results in others when the R^2 scoring metric was used. Another important observation is that in none of average performance metrics does R^2 optimisation ever perform better than a negative value. This means that on average the model does not capture much of the variability of the data. This is true even when the performance of the model is good in the case of the Random Forest with MAE scoring and 6 features in the city of San Juan. This could be because the coefficient of determination score is too sensitive to major outliers. It is sometimes able to predict when an outbreak (greater than 100 cases for example) will occur though not necessarily accurately how much. This skews the performance metrics and may result in the discarding of promising models in the results. On the other hand, the good performance of the MAE scoring metric might be due to the fact that it does not overly penalise outliers like RMSE (Root mean square error) or coefficient of determination. However, this does not mean we should not use the coefficient of determination at all as a performance metric as it can still provide us useful information when used to compare two different model's ability to explain variability in the data, this is evidenced as the Support Vector Machine with the MAE scoring metric for the hyperparameter optimization actually had a better R^2 performance metric than when R^2 was itself used as a scoring metric for the hyperparameter optimization showing that the former model explained the variability in the model better.

A comment should be made on the city of Iquitos where varying the scoring metric for the hyperparameter optimization did not produce any statistically significant changes in performance metrics of any of the 3 models, it again failed to predict any outbreaks. In addition to the aforementioned reason that the dataset might be too small, a separate and more thorough and specific feature importance analysis and feature engineering potentially via principal component analysis may need to be constructed to improve the results for the city.

Finally, when we increased the number of features in the model from 6 to 10 the model did not show any increased performance for both cities and in fact had worse performance on all performance metrics. It also failed to predict certain outbreaks as seen in figure 4 and its sensitivity to outbreaks was noticeably diminished. They also show a greatly diminished ability to predict major dengue outbreaks. This underperformance could be due to the fact that adding unnecessary features (that was removed through the lasso feature importance analysis) may make it harder to understand the complex relationships between the target variables and the underlying features. There could be muti-collinearity and much more data would be needed to decipher the underlying relationships between the variables. This can often lead to overfitting and poor generalisation of results to testing data. One technique to address these issues is to use ensemble methods like Random Forest that aggregate predictions from multiple base models, but in this case even Random Forest was not enough to prevent the curse of dimensionality though it did improve sensitivity to outbreaks.

V. Conclusion and Recommendations

The paper successfully analysed the time series dataset and created a powerful model that can forecast dengue outbreaks and potentially predict dengue cases with a reasonable level of error/performance. It investigated the result of changing the scoring metric in the hyperparameter optimization phase of building the model. The results suggested that in the city of San Juan using the mean absolute error as a scoring metric instead of the coefficient determination in the Bayesian hyperparameter optimization allowed a significantly better performance metrics for the Support Vector Regression and greatly increased the ability of Random Forest models to predict Dengue outbreaks. The results for the city of Iquitos were however poor where none of the models were able to predict Dengue outbreaks accurately.

The results did not lead to many conclusive results. Changing the scoring metric and increasing the number of the features did not improve the performance metrics or the model's ability to predict outbreaks for the city of Iquitos. We cannot be sure that models for Iquitos will be effective in predicting any rare events like Dengue outbreaks and as recommended above a separate feature importance analysis may need to be conducted for the city and further feature engineering techniques may need to be implemented to better enhance the ability of the models for the city.

In the city of San Juan, increasing the number of features however diminished and worsened the performance of the support vector regression and also decreased the ability of the Random Forest model to predict Dengue outbreaks. Further studies should examine the effects of specific variables on Dengue spread more accurately to potentially get more important features and help enhance the results obtained in this study. Performance of models seem to be dependent on the city and amount of data, this is corroborated by this study. Another very important area of research is to experiment on the use of different scoring metrics and potentially hybrid scoring metrics that take into account different performance metrics like MAE, RMSE and R^2 together in different ratios (or in more complicated ways) and experiment on their effects on the predictive ability of the models.

REFERENCES

[1] S. Anno *et al.*, "Spatiotemporal Dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning," *Geospatial Health*, vol. 14, no. 2, Nov. 2019. doi:10.4081/gh.2019.771

[2] DrivenData, "Dengai: Predicting disease spread," DrivenData, https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/ (accessed Apr. 29, 2024).

[3] C.-Y. Kuo, W.-W. Yang, and E. C.-Y. Su, "Improving dengue fever predictions in Taiwan based on feature selection and random forests," *BMC Infectious Diseases*, vol. 24, no. S2, Mar. 2024. doi:10.1186/s12879-024-09220-4

[4] Chathika Gunaratne, undefined., et al, "Evaluation of Zika Vector Control Strategies Using Agent-Based Modeling," 2016.

[5] U. Khaira, P. E. Utomo, R. Aryani, and I. Weni, "A comparison of sarima and LSTM in forecasting dengue hemorrhagic fever incidence in Jambi, Indonesia," *Journal of Physics: Conference Series*, vol. 1566, no. 1, p. 012054, Jun. 2020. doi:10.1088/1742-6596/1566/1/012054

[6] S. K. Dey *et al.*, "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach," *PLOS ONE*, vol. 17, no. 7, Jul. 2022. doi:10.1371/journal.pone.0270933

[7] K. Roster, C. Connaughton, and F. A. Rodrigues, "Machine-learning–based forecasting of dengue fever in Brazilian cities using epidemiologic and meteorological variables," *American Journal of Epidemiology*, vol. 191, no. 10, pp. 1803–1812, May 2022. doi:10.1093/aje/kwac090

[8] S. Patil and S. Pandya, "Forecasting dengue hotspots associated with variation in meteorological parameters using regression and time series models," *Frontiers in Public Health*, vol. 9, Nov. 2021. doi:10.3389/fpubh.2021.798034

[9] C. M. Benedum, K. M. Shea, H. E. Jenkins, L. Y. Kim, and N. Markuzon, "Weekly dengue forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore," *PLOS Neglected Tropical Diseases*, vol. 14, no. 10, Oct. 2020. doi:10.1371/journal.pntd.0008710

[10] J. Ranstam and J. A. Cook, "Lasso regression," *British Journal of Surgery*, vol. 105, no. 10, pp. 1348–1348, Aug. 2018. doi:10.1002/bjs.10895

[11] "What is lasso regression?," IBM, https://www.ibm.com/topics/lasso-regression (accessed Apr. 29, 2024).

[12] K. Ito and R. Nakano, "Optimizing support vector regression hyperparameters based on cross-validation," *Proceedings of the International Joint Conference on Neural Networks, 2003.*, 2003. doi:10.1109/ijcnn.2003.1223728

[13] "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," in IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 781-781, May 2005, doi: 10.1109/TNN.2005.848998.

[14] G. L. Team, "A complete understanding of lasso regression," Great Learning Blog: Free Resources what Matters to shape your Career!, https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Linear%20regression%20model%3A%20LASSO%20regression (accessed Apr. 29, 2024).

[15] R. Hossain and D. Timmer, "Machine Learning Model Optimization with Hyper Parameter Tuning Approach," *Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence*, vol. 21, no. 2, 2021.

[16] J. Brownlee, "Nested cross-validation for Machine Learning with python,"MachineLearningMastery.com, https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/ (accessed Apr. 29, 2024).

[17] J. Wu *et al.*, "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019. doi:https://doi.org/10.11989/JEST.1674-862X.80904120

[18] V. C. Paraná, C. A. Feitosa, G. C. da Silva, L. L. Gois, and L. A. Santos, "Risk factors associated with severe dengue in Latin America: A systematic review and meta-analysis," *Tropical Medicine & International Health*, vol. 29, no. 3, pp. 173–191, Jan. 2024. doi:10.1111/tmi.13968

VI. CONTRIBUTIONS

Muhammad Hassan Rizvi: Conceptualisation, Methodology, Software, Validation, Investigation, Writing – Original Draft. **Deniz Sagmanli:** Conceptualisation, Methodology, Software, Validation, Investigation, Writing – Original Draft, Writing – Review & Editing. **Sundeep Veluchamy**: Conceptualisation, Methodology, Software, Validation, Investigation, Writing – Original Draft, Visualisation.