STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, April 2026, pp 3320–3340.
Published online in International Academic Press (www.IAPress.org)

Machine Learning Algorithms for the Prediction of the Spread of COVID-19: A Comparative Analysis using Data from Namibia

Claris Shoko [1,*], Eriyoti Chikodza [2]

[1]*Department of Statistics, University of Botswana, Plot 4775 Notwane Rd Gaborone Botswana*
[2]*Department of Mathematics, University of Botswana, Plot 4775 Notwane Rd Gaborone Botswana*

Abstract
Background and Objective: Robust prediction of daily COVID-19 transmission dynamics is important for effective public health management. This paper examines, evaluates and compares the predictive capabilities of selected machine-learning algorithms in describing the spatial and temporal spread of COVID-19 in Namibia. **Materials and Methods:** In this study, four machine-learning models, namely the Support Vector Machine (SVM), the TBATS model, the Generalized Additive Model (GAM), and the Stochastic Gradient Boosting Machine (SGBM), were calibrated to daily COVID-19 data. Initially, the performance of each model was evaluated visually using forecast plots constructed from the test dataset. Subsequently, a quantitative comparison was implemented using key performance indicators (KPIs) that are not sensitive to outliers, namely mean absolute percentage error (MAPE) and the mean absolute error (MAE). **Results:** The analysis of results demonstrated that the positive rate, reproductive rate, and stringency index are statistically significant determinants of COVID-19 spread in Namibia, with all corresponding p-values below 0.05. Considering the models involved in the study, the machine learning approaches, in particular, the GAM, SGBM and SVM radial kernel were among the top 3 superior long-term forecasting models, whereas the traditional approach, the TBATS, was more robust on short-term forecasting, followed by the SVM linear and the GAM, in that order. Combining machine learning models improves forecasting performance in the long run. **Conclusion:** The research results propose the SVR(Radial)-GAM-SGBM, a model that combines the SVR radial kernel, GAM and SGBM as appropriate models for long-term prediction and for identifying key regressors influencing COVID-19 transmission dynamics. Effective short-term forecasts provide critical early-warning mechanisms for the health sector, enhancing intervention plans. These upgraded forecasting methodologies support national and global endeavours to achieve Sustainable Development Goal 3, which focuses on ensuring health and well-being for all. **Implications:** Integrating robust machine-learning models into public health management systems can strengthen epidemic surveillance and control by enabling effective and timely interventions, rational deployment of national healthcare resources, and data-supported crafting

Keywords COVID-19; Machine learning algorithms; SVM kernel functions; Generalised additive model; Stochastic gradient boosting machine; TBATS; Namibia.

AMS 2010 subject classifications 68T05, 62M20

DOI: 10.19139/soic-2310-5070-2698

1. Introduction

The history of humanity has been characterised by outbreaks of epidemics that claimed millions of lives. Some documented examples of these human adversaries include the Black Death, Spanish flu, Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) [1]. In recent years, 31 December 2019 heralded the onset of yet another difficult page for the international community, with several pneumonia cases

*Correspondence to: Claris Shoko (shokoc@ub.ac.bw). Department of Statistics, University of Botswana, 4775 Old Notwaane Rd, Gaborone, Botswana.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

of unknown origin exploding in Wuhan, Hubei Province, China [2]. Scientific investigations, including laboratory deep sequencing analysis of lower respiratory tract samples from diverse patients, showed that the novel severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) causes this illness. As time progressed, the disease was code-named COVID-19 by the World Health Organisation (WHO). The sub-discipline of Microbiology classifies these coronaviruses as members of the subfamily Coronavirinae within the Coronaviridae family and the order Nidovirales. As COVID-19 continued to ravage communities across the globe, the WHO declared the highly infectious and deadly disease a pandemic of global proportions on 11 March, 2020 [3]. After some careful study of the dynamics of COVID-19, the initial cluster was identified to be in Wuhan's major seafood market. Having realised that COVID-19 had rapidly presented itself as a formidable socio-economic threat to the human race, the WHO promulgated a global epidemiological alert close to the end of April 2020. By 23 February 2023, close to 674-million COVID-19 cases had been confirmed and documented, while COVID-19-related deaths in the vicinity of 6.87-million were recorded [4]. The COVID-19 scourge persists to haunt the communities as the global cumulative total continued to rise at a rate in the range of 163,000 to 293,000 confirmed daily cases at the height of the epidemic between 2020 and 2021 [4].

COVID-19 disrupted the global socio-economic ecosystems, and this prompted nations and multilateral establishments, such as the WHO, to design strategies for combating the virus. For instance, [5] analysed how COVID-19 negatively affected access to sun-and-beach tourism in the Balearic Islands, thereby maiming one of the region's key streams of national revenue. A possible avenue for decelerating the spread of the virus and consequently mitigating its impact on the global economy was to control the physical interactions of people [6]. This was achieved by imposing temporal closure of educational institutions, restricting the movement of people from one geographical locality to another, and the cancellation or postponement of events. Unfortunately, such measures have various undesirable social, economic, and psychological consequences [6]. Not unexpectedly, authorities and policymakers worked towards striking the delicate balance between curtailing the spread of the virus and lessening the impact of these psycho-socio-economic ramifications. In order to craft and implement effective COVID-19-related policy, leadership relies on scientific research into the dynamics of the disease both at the cellular and population levels [7].

The fragile developing African economies have had their fair share of the adverse impact of the COVID-19 pandemic. Despite the fact that Africa was affected at a relatively much later stage of the COVID-19 pandemic evolution, researchers and other experts are convinced that the impact of the scourge was most disastrous on the continent [8], [9] and [10]. Recent reports from authoritative sources indicate that although there has been notable progress in fighting COVID-19 in the Southern African Development Community (SADC) sub-region, the war is not over, hence the need for continued research into the diverse aspects of the disease [11].

Prediction of the rate of COVID-19 spread and modelling of its course have a critical impact on both the health system and policy makers [12]. Indeed, policy making depends on judgments formed by the prediction models to propose new strategies and to measure the efficiency of the imposed policies. Based on the nonlinear and complex nature of the disorder emanating from COVID-19 and difficulties in estimation of virus transmission features, various traditional and modern methodologies, and techniques have been applied to predicting the spreading trajectory of the pandemic [12], [13]. Some of the widely applied instruments for analysis and forecasting purposes are anchored in at least one of the sub-disciplines of statistical modelling, mathematical epidemiology, artificial intelligence, and machine learning [3], [14], [15], [8], [12] and [7]. Other prominent approaches used to enhance prediction results encompass adaptive neuro-fuzzy inference systems and recurrent neural networks.

Research has amply demonstrated that the spread of COVID-19 is random and is influenced by many factors [12]. For this reason, robust forecasting models cannot be achieved if only a single paradigm or tool is adopted for developing a prediction architecture. The non-linearity and unstructured nature of COVID-19 data have necessitated the adoption of computational intelligence [13], artificial intelligence, and machine learning methods in fighting the pandemic. For instance, fractal dimension and interval type-3 fuzzy logic [16], multiple linear regression, machine learning [17], multilayer perceptron, grey prediction model, simulation model, Holt model, LSTM model, and support vector machine, have significantly improved the prediction of COVID-19 behaviour [18].

This study focuses on short-term prediction of COVID-19 in Namibia, a southern African nation with a population of 2.5 million. According to [19] is classified as a high to middle-income country with an impressive life expectancy of 64, an adult literacy index of 91%, and high rates of primary school enrolment of 97%. On 13 March 2020, the first two cases of COVID-19 were reported in the Namibian capital, Windhoek, when a Romanian couple landed in the country [20]. In response to that report, the Namibian Minister of Health convened a meeting of the National Health Emergency Management Committee Special Committee on COVID-19 on 14 March 2020. The outcome of the meeting was the declaration of a national health emergency. Within a short space of time, the viral disease invaded all fourteen (14) regions of the country. Borrowing a leaf of experience from various other jurisdictions across the world, in response to reports of early confirmed COVID-19, the Namibian authorities implemented a regime of non-pharmaceutical interventions (NPIs) on a gradual scale. Such interventions encompassed isolation of confirmed or suspected cases, contact tracing, closure of educational and other training establishments, and the prohibition of mass gatherings [8]. With effect from 24 March 2020, a ban on entries into the country was imposed, and returning citizens had to comply with a 14-day self-quarantine. A further identification of 11 COVID-19 cases triggered the implementation of tight measures, leading to a COVID-19-induced lockdown in the regions of Khomas and Erongo. A which was initially introduced on 27 March 2020, was stretched by an additional two weeks to envelop the whole country. The lockdown was lifted on 4 May 2020.

The thrust of this study is to compare several different ensemble machine-learning techniques as applied to COVID-19 prediction. The approaches examined include the support vector machine (SVM), the TBATS model, the generalised additive model (GAM), and the Stochastic Gradient Boosting Machine (SGBM) approach are used as competing machine learning models. Comparison of the fitted machine learning methods is done using the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

1.1. Contributions

This research made six major contributions to the existing literature, which are explained as follows.

- Novel ranking and benchmarking of different machine-learning models for prediction of epidemic dynamics. The research examines four different forecasting approaches, namely, SVM, TBATS, GAM, and SGBM, using daily COVID-19 data from a developing country. On the basis of initial visual inspection complemented by rigorous KPI-based evaluation, the paper develops a benchmarking framework specifically designed for nonlinear epidemiological time series analysis.
- Incorporation of spatial–temporal and policy-related factors into predictive time-series modelling. Through incorporating the positive rate, reproductive rate, stringency index, and vaccination-related variables as statistically significant predictors of COVID–19 dynamics, the research established an enhanced time series modelling paradigm. The integration of policy intervention measures further buttresses the interpretability of epidemic forecasting methods.
- Data-based ranking of models in nonlinear epidemic time series analysis. By using empirical data, the GAM and linear-kernel SVM proved to be the best-performing models, illustrating the criticality of smooth functional estimation, and linear decision boundaries for short-term forecasts. The high forecasting performance of these two models justifies the use of models that capture nonlinearities in health-related data.
- Establishment of a reliable scientific early-warning forecasting mechanism for policy makers in the public health sector. The research provided validation of models on a rolling forecasting window (14-, 30-, and 60-day horizons) and aligned model outputs with real epidemiological waves. This approach yielded an easy-to-use early-warning mechanism for the public health sector. This early-warning framework strengthens preparedness by facilitating timely interventions and optimal deployment of resources by health authorities.
- Alignment of machine-learning forecasting machinery with global health objectives. The paper contributes to the achievement of Sustainable Development Goal 3 by illustrating how contemporary technology-based forecasting models can support disease surveillance, optimise control strategies, and improve health-system resilience at the national level.
- Proposal and application of smoothed time-series normalisation technique and an ensemble forecasting model. To handle the non-normality and zero-inflation of COVID-19 data, the research introduced and applied a spline-based smoothing approach ($\lambda = 0.472716$) that fixes technical complications associated with

log-transformation. The research demonstrates how model ensembling and smoothing techniques can address the limitations of noisy epidemiological time series.

1.2. Organization

The remainder of this paper is structured as follows. Section 2 provides a detailed exposition of the methodological framework employed in the study. Section 3 presents the empirical results derived from the modelling procedures, while Section 4 offers a critical discussion and interpretation of these findings. Finally, Section 5 concludes the paper by synthesising the key insights and outlining recommendations for future research.

2. Materials and Methods

2.1. Data

In this study, we used an openly available daily number of confirmed cases of COVID-19 reported by Our World in Data (https://www.ourworldindata/coronavirus-source-data) from 14 March 2020 to 20 June 2022.

Figure 1 provides a schematic summary of the analysis procedure for predicting the spread of COVID-19 in Namibia.

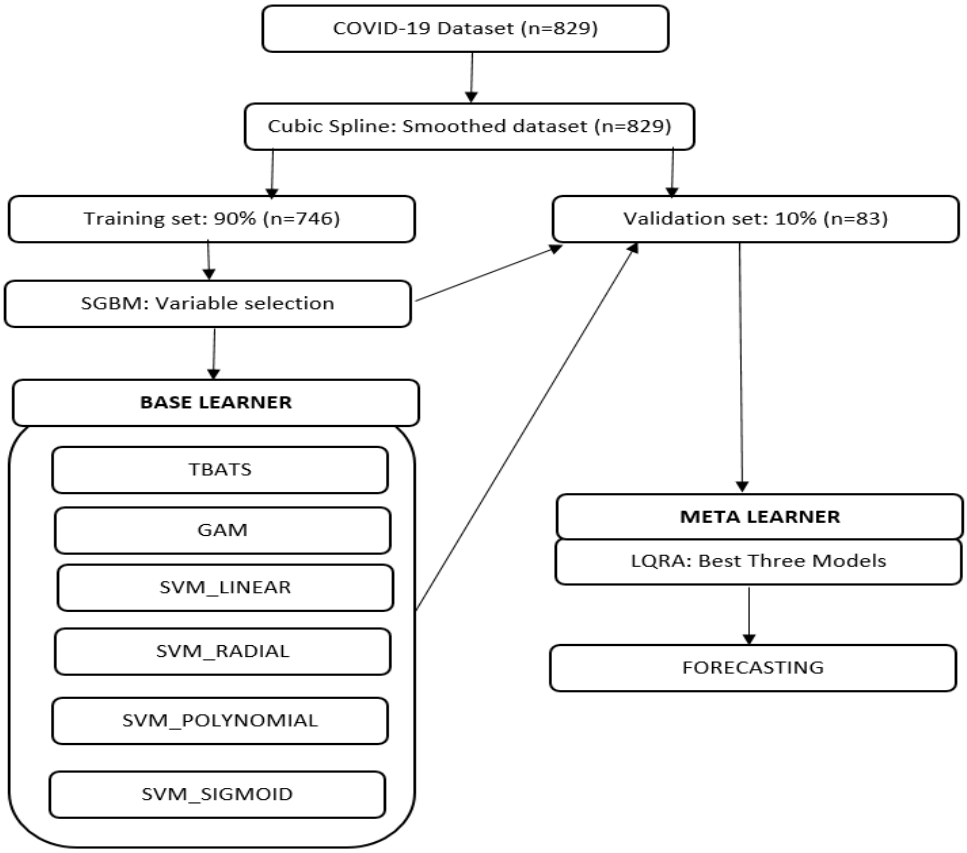

Figure 1. Schema for the machine learning algorithms for predicting COVID-19 in Namibia.

2.2. Machine Learning Algorithms

2.2.1. Trigonometric seasonality Box-Cox transformation ARIMA errors Trend Seasonal components (TBATS) model The TBATS model uses the Box-Cox transformation, exponential smoothing, trigonometric seasonality and ARMA errors. It is generally used for forecasting time series with complex seasonal patterns. The components of the model are:

- The Box-Cox transformation

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{(\omega)}}{\omega}; & if \ \ \omega \neq 0 \\ \log y_t; & if \ \ \omega = 0 \end{cases} \tag{1}$$

  where y_t is the confirmed daily cases on day t, ω is the transformation parameter, and log denotes the natural logarithm.
- Deterministic and stochastic trend

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-1}^{(i)} + d_t, \tag{2}$$

$$l_t = l_{t-1} + \phi b_t + \alpha d_t, \tag{3}$$

$$b_t = (1-\alpha)b + \alpha b_{t-1} + \beta d_t \tag{4}$$

  where T denotes the number of seasonal patterns, l_t is the local trend in period t, b represents the long-run trend, b_t denotes the short-run trend in period t, $s_{t-1}^{(i)}$ represents the i^{th} seasonal component at time $t-1$, d_t denotes the ARMA(p,q) process and α, β and Φ are smoothing parameters.
- Trigonometric seasonality

$$s_t^{(i)} = \sum_{j=1}^{k_1} s_{j,t}^{(i)}, \tag{5}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t, \tag{6}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t, \tag{7}$$

  where $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$ are smoothing parameters and $\lambda_{(j)}^{(i)} = \frac{2\pi_j}{m_i}$ with m_i representing the period of the seasonal cycle.
- ARMA errors

$$d_t = \sum_{i=1}^{p} \phi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t, \tag{8}$$

  where ϕ_i, θ_i denote the autoregressive and moving average parameters, respectively and ϵ_t is a white noise process. The 4 components put together give the TBATS model.

2.2.2. Gradient Boosting Method (GBM) Gradient boosting (GB) is a machine learning technique that fits an additive model in a stage-wise way. The additive model can take the form given in Equation (9).

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m), \tag{9}$$

where $b(x; \gamma_m) \in \Re$ are functions of x which are characterised by the expansion parameters β_m, γ_m. The parameters β_m and γ_m are fitted in a stage-wise way, a process which slows down over-fitting [21]. Stochastic gradient boosting (SGB) is an extension of GB in which a random sample of the training data set is taken without replacement. See Friedman for a detailed discussion of the gradient boosting method [22]. The SGBM is very good at ranking explanatory variables according to their relative influence on the response variable. In this study, variables selected by the SGBM are used to fit the generalised additive model (GAM) and the support vector machine models (SVMs)

2.2.3. Generalized additive models The generalised additive model (GAM) is a machine learning algorithm that fits smooth relationships between variables with complex relationships that can not be handled by standard linear and nonlinear models [23]. Let y_t denote the confirmed daily cases on day $t, t = 1, \ldots, n$ with the corresponding covariates $x_{t1}, x_{t2}, \ldots, x_{tp}$, where p represents the number of variables. The generalised additive model is then written as:

$$y_t = \beta_0 + \sum_{j=1}^{p} s_j(x_{tj}) + \epsilon_t, \tag{10}$$

where β_0 is a constant parameter, s_j are smooth functions and ϵ_t are independent and identically distributed (i.i.d) error terms. Equation (11) is estimated using penalised cubic splines given as:

$$\min_{s_j} \left[\sum_{t=1}^{n} \left(y_t - \beta_0 - \sum_{j=1}^{p} s_j(x_{tj}) \right)^2 + \sum_{j=1}^{p} \lambda_j \left(\int \left(f''(x) \right)^2 dx \right) \right], \tag{11}$$

The penalty parameter controls the degree of smoothness $\Lambda = (\lambda_j, j = 1, \ldots, p)$ which is optimised using the generalised cross-validation criterion (GCV) [**?**]. The smooth function, s_j, is a sum of basis functions, $b_i(x)$, together with their regression coefficients β_i and is given by $s_j(x) = \sum_{j=1}^{p} b_i(x)\beta_i$, where q denotes the basis dimension.

2.2.4. Support vector regression Support vector machine (SVM) is a powerful machine-learning algorithm for recognising complex patterns [24] such as the one characterised by the daily COVID-19 cases. When SVM is used in regression modelling, it is referred to as support vector regression (SVR). For SVR, the kernel functions transform the data into higher higher-dimensional feature space to make it possible to perform linear separation. Linear separable cases make it possible to find the hyperplane by maximising the classifier margin. When input data are nonlinear separable, SVMs can produce accurate and robust classification results on a sound theoretical basis. This can be done by means of kernel transformation. The kernel method enables modelling of higher-dimensional, non-linear problems by adding dimension to the raw data and thus transforming it into a linear problem. The mathematical setup of SVMs is as follows:

Let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ be the training set where $y_i \in (+1, -1), i = 1, 2, \ldots, l$, and $x_i \in \Re^n$. The optimisation is as follows:

$$\min_{(\omega,b,\xi^+,\xi^-)^2} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i^+ + \xi_i^-), \tag{12}$$

subject to:

$$\begin{cases} \langle \omega, x_i \rangle + b - y_i - \varepsilon \leq \xi^+, & i = 1, 2, \ldots, l \\ y_i - \langle \omega, x_i \rangle + b - \varepsilon \leq \xi^-, & i = 1, 2, \ldots, l \\ \xi^+, \xi^- \geq 0, & i = 1, 2, \ldots, l \end{cases}$$

Where ω, b are the parameters of the hyperplane, C is the regularisation parameter introduced to measure the trade-off between the complexity and loss, ξ^+, ξ^- are the slack variables one for exceeding the target by more than ϵ and the other below the target value helping to allow consideration of badly classified points. When the kernels are introduced, the optimisation problem becomes:

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{l} y_i\alpha_i - \varepsilon\sum_{i,j=1}^{l}|\alpha_i| - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j K(x_i, x_j) \\ \text{s.t.} \quad & \sum_{i=1}^{l}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, l \end{aligned}$$

Where α_i are the Lagrange multipliers and $K(x_i, x_j)$ represents the kernel function. For this study, we used the linear, radial basis, polynomial, and sigmoid kernel functions as outlined in Table 1 below.

Table 1. kernel functions.

Kernel function	Formula	Optimization parameter
Linear Kernel	$K(x_i, x_j) = x_i^T.x_j$	C, γ
Radial Kernel	$K(x_i, x_j) = exp\left(-\gamma\|\|x_i - x_j\|\|^2 + C\right)$	C, γ
Polynomial Kernel	$K(x_i, x_j) = \left(x_i^T.x_j + r\right)^d$	C, γ, r, d
Sigmoid Kernel	$K(x_i, x_j) = tanh\left(\gamma x_i^T.x_j^+ r\right)$	C, γ, r
C= cost, $\gamma = gamma$, $r =$ coefficient, $d =$ degree		

γ is a kernel-specific parameter for non-linear kernel functions that defines how far the influence of a single support vector reaches. γ takes any value in the interval $(0, 1)$. The whole idea behind this is to determine the result of the function for predicting the sample

$$f(x) = sign\left[\sum_{i=1}^{l} y_i\alpha_i K(x_i, x_j) + b\right], \tag{13}$$

Where x_i and y_i are the support vectors and their membership classification, respectively. The performance of the SVM depends on the choice of the kernel function. Hence, it is always recommended to use as many kernel functions as possible and then select the most accurate kernel function.

SVR analysis is done using the "e1071" package for R software. The package uses the function "svm()" that automatically switches to regression mode using the epsilon function (epi-function) if y is a numeric variable. The

function, by default, internally scales the input features and the response variable to zero mean and unit variance for better performance; these scaling parameters are stored for future predictions. The "e1071" package includes a tune function that allows for hyperparameter selection via cross-validation and grid-search over a specified range of values of the cost parameter and γ.

2.3. Combined forecast

Forecasts from the best three performing models are combined using the linear quantile regression averaging (LQRA) approach. LQRA is a method of combining multiple forecasts into a single probabilistic forecast by using the forecasts as inputs to a quantile regression model. Thus, it allows for the integration of various forecasting models without needing to develop a new, complex model from scratch. According to Shoko and Sigauge [25] we let $y_{t,k,\tau}$ be the smoothed daily COVID-19 forecasts for the next 84 days, with $k = 1, 2, 3$ total number of best performing methods for predicting the next observation, $\tau \in (0, 1)$. The combined forecasts are given by:

$$\widehat{y}_{i,t,\tau}^{\text{LQRA}} = \beta_{0,t,\tau} + \sum_{k=1}^{3} \beta_k \widehat{y}_{t,k,\tau} + \epsilon_{t,\tau}. \tag{14}$$

.

2.4. Variable Selection

Gradient Boosting approach: variables are selected using the gradient boosting approach. This approach has an in-built mechanism for selecting variables that contribute to the variable of interest (response variable). The selected variables are ranked in order of their relative influence. The selected variables will be used in fitting the SVR and GAM models. The SVR model will use the linear, radial (Gaussian), polynomial, and sigmoid kernels.

2.5. Key Performance Indicators (KPIs)

The key performance indicators (KPIs) are given according to [**?**] as follows:

Mean absolute percentage error (MAPE)- most commonly used KPIs to measure forecast accuracy, and is given by

$$MAPE = \frac{1}{n} \sum |\frac{e_t}{y_t}|, \tag{15}$$

However, MAPE is a poor-accuracy indicator.

Mean absolute error (MAE) is less sensitive to outliers compared to the root mean square error (RMSE) and the mean square error (MSE). Thus, making it more robust nonlinear time series data that have high variability or extreme values, such as the COVID-19 time series data. This makes the MAE a very good KPI to measure forecast accuracy in this study. The formula for MAE is given by

$$MAE = \frac{1}{n} \sum |e_t|, \tag{16}$$

Root mean squared error (RMSE)- a strange but very helpful KPI given by

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{m} (e_t)^2} \tag{17}$$

where $e_t = y_t - \hat{y}_t$. However, RMSE is sensitive to outliers. Therefore, although many KPIs will be calculated in this study, we mainly focus on using the MAE in selecting the best-performing model since the time series dataset is positively skewed.

3. Results

The time series data for daily COVID-19 cases considered in this study is made up of 829 observations, that is, from 14 March 2020 to 20 June 2022. For this period, the minimum number of daily COVID-19 cases recorded is zero, and the maximum is 4210. The average number of reported cases is 203.7 per day, with a median of 84 cases. Thus, the median is far too low compared to the mean, an indication of non-normality in the distribution of the data.

3.1. Exploratory data analysis

In Table 2 below we present the descriptive statistics for some of the important variables in the study. These variables include total COVID-19 cases (TC), new reported cases (NC), total deaths (TD), total vaccines (TV), people fully vaccinated (PFV), new vaccines (NV), and stringent index (SI).

Table 2. Descriptive statistics for some important variables in the study.

	TC	NC	TD	TV	PFV	NV	SI
Min	2	0	0	0	0	0	8.33
1st Qu	11714	13	126	0	0	0	28.7
Median	48654	83	643	0	0	0	42.59
Mean	72716	203.7	1701	31598	13248	283.1	44.57
3rd Qu	129133	214	3572	0	0	0	54.63
Max.	168904	4210	4056	825518	426681	9823	87.04

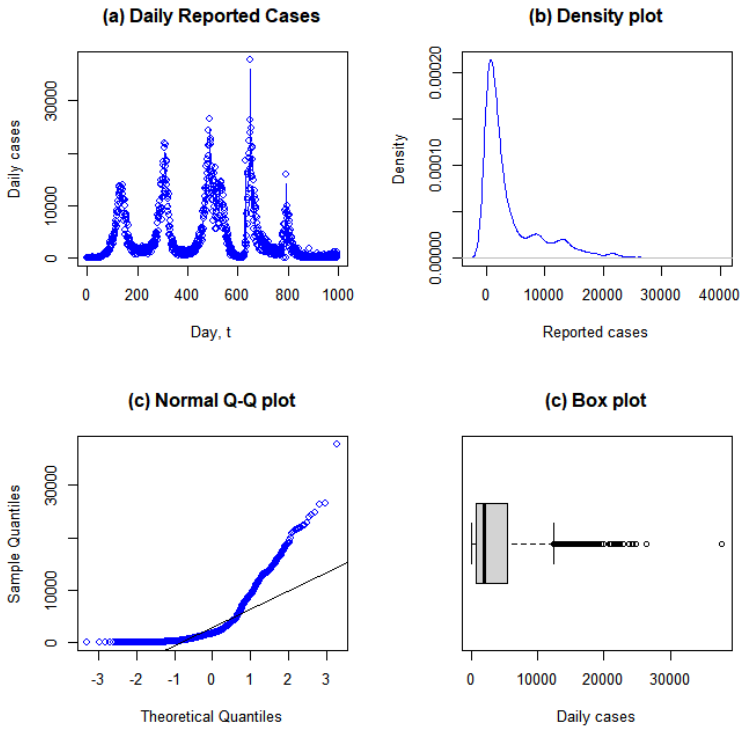

Figure 2. Univariate data analysis.

Results presented in Table 2 show that a maximum of 168,904 cases were reported in Namibia by June 2022. For the period under study, an average of approximately 204 cases were reported daily, with a maximum number of cases of 4,210. By June 2022, 825,518 total vaccines had been rolled out in Namibia. The country recorded a total of 4,056 deaths by the end of the period under study.

A univariate exploratory data analysis for daily COVID-19 cases is presented in Figure 2. Figure 2 presents the time series plot for the daily COVID-19 cases, the normal Q-Q plot, the density plot, and the Box plot. The Box plot, density plot and the Q-Q plot confirm that the daily COVID-19 cases are not normally distributed.

Several approaches can be used to handle non-normal time series data, including log-transforming. However, for this dataset, the log-transformation is not appropriate because of the zero reported cases on some days. Thus, the logarithm of zero is undefined. We transformed the dataset by smoothing the original series using the cubic regression spline approach. The smoothed time series removes the random noise and short-term fluctuations, making it easier to identify and analyse the underlying trends, seasonality, and patterns. Smoothing helps improve the clarity of data for visualisation and forecasting. A smoothing parameter $\lambda = 0.472716$ is used. For analysis, the smoothed time series is coded as "nfit", a fitted non-linear trend of the series. The smoothed time series together with plots of the residuals as well as the distribution of the residuals are presented in Figure 3. The density plot from the residuals suggests a normal distribution.

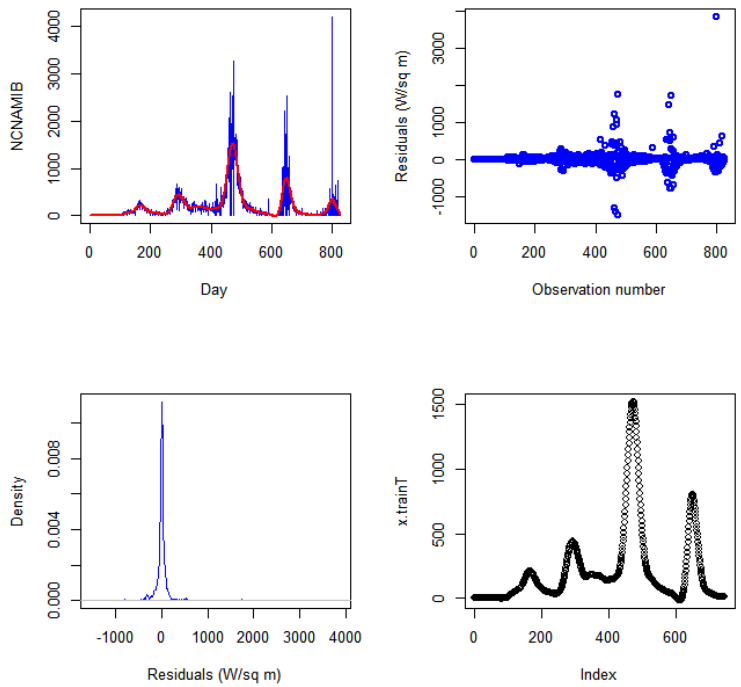

Figure 3. Smoothed COVID-19 cases based on the Cubic Spline approach.

The dataset consists of 36 variables that include total cases (TC), new cases (NC), total deaths (TD), new deaths (ND), reproductive rate (RR), total tests (TT), new tests (NT), positivity rate (PR), people fully vaccinated (PFV), new vaccines (NV), stringent index (SI), to mention a few. As presented in Figure 4 below, the data set did not have any missing values since the zero values indicate no reported COVID-19 cases, especially during the first stages and the last stages of the pandemic.

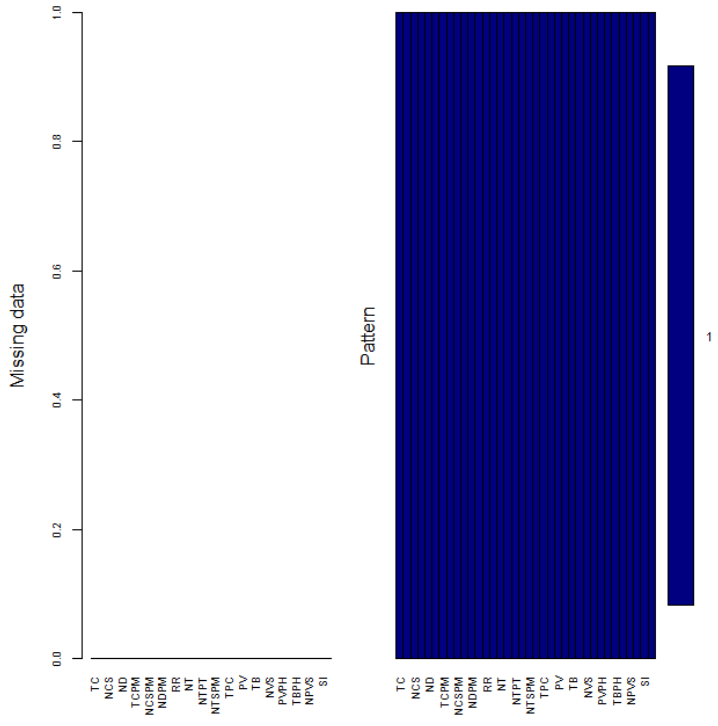

Figure 4. Exploring missing data.

A multivariate analysis of the variables is done by first presenting a matrix grid for the correlation of variables, which is shown in Figure 5 below.

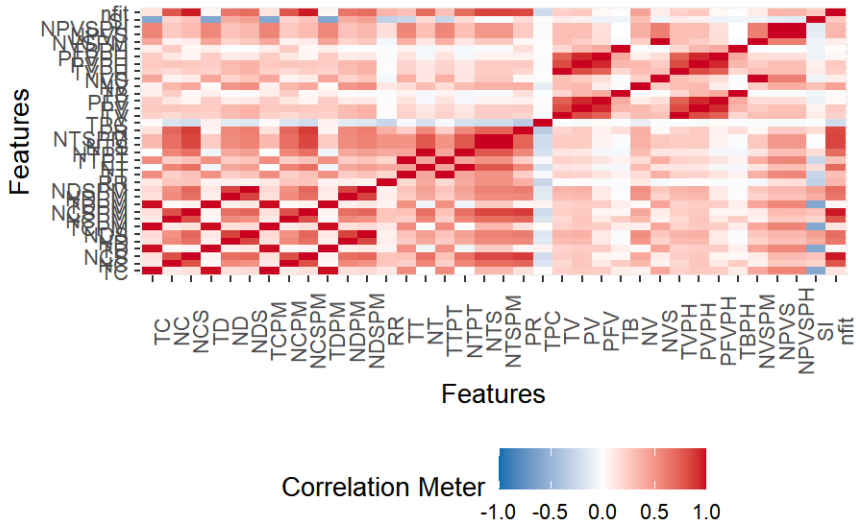

Figure 5. correlation grid matrix. for the variables.

For this study, the response variables of interest are the smoothed daily COVID-19 cases (nfit). Figure 5 shows that nfit is not correlated with TD, TC, total cases per million (TCPM), total deaths per million (TDPM), PV, PFV, SI, and new vaccines smoothed (NVS). These are some of the covariates that will be considered when fitting the machine learning models.

3.2. Application of machine learning algorithms on the smoothed daily COVID-19 cases in Namibia

The data has 829 observations. 90% (746) of the observations are used as the training set, and the remaining 10% (84) represent the test dataset. The choice of the 10% test set for validation is based on the fact that this was the beginning of the last wave of the COVID-19 pandemic. Thus, we are interested in testing whether the fitted models can predict this wave based on their knowledge of the previous waves. However, further validation of the fitted models is done on 60- (2 months), 30- (1 month), and 14-day (2 weeks) rolling windows. These windows are used for the validation of the performance of fitted models. Several candidate machine learning models are considered in this section. These include the SGBM, GAM, and SVM models benchmarked on the TBATS model. A robust comparison of the performance indicators (ME, RMSE, MAE) for the fitted models is done. Based on this comparison, the best 3 models for each rolling window are selected. In addition, plots of the performance of fitted models on the validation set are compared.

3.2.1. TBATS model: The TBATS model is fitted as a benchmark for the performance of the machine learning models. From the 746-day training set a TBATS(1, {3,2}, 0.906, -) was proposed. The fitted model does not show any seasonal variation, represented by the - sign, because the spread was mainly impacted by human behaviour and public health interventions, thereby distorting natural transmission patterns. One would expect the highest number of cases to be recorded in winter, but this was not the case since stringent measures were high, especially around winter, until a greater percentage of susceptible individuals got vaccinated. Thus, the Box-Cox transformation $\omega = 1$ indicates no transformation was applied (this is possible since the data has already been smoothed), the damping parameter = 0.906, which is less than 1, meaning the forecast will gradually flatten out to a constant slope. However, the trend close to 1 implies a slow/gradual decaying trend, autoregressive order p=3 represented by the coefficients $p_1 = 1.684957, p_2 = -0.449888, p_3 = -0.246601$ and a moving average of order 2 with $q_1 = -0.776896, q_2 = 0.157973$. The TBATS model did not find any seasonal period in the smoothed time series dataset. The model has a mean error of 132.73, root mean square error of 491.13, and mean absolute error of 134.91. However, the disadvantage of the TBATS model is that it is a purely data-driven time series model and generally does not inherently incorporate epidemiological factors such as transmission rates, or the impact of non-pharmaceutical interventions such as social distancing and lockdowns. For a more robust and long-term forecasting, we propose machine learning approaches that incorporate external factors and epidemiological contexts that influence the spread of COVID-19. Machine learning approaches to model and forecast the spread of COVID-19 are presented in the subsections that follow. Further performance of the TBATS on the other rolling windows is presented in Section 3.3

3.2.2. SGBM: The SGBM model helps to identify the most significant factors influencing COVID-19 transmission and severity. In fitting the stochastic gradient boosting machine (SGBM) model, we consider the variables selected from the correlation grid, that is, variables that either have no correlation or are not highly correlated with the response variable of interest, which is the smoothed COVID-19 cases. For the SGBM model, resampling was done using bootstrapping with 25 replications. 21 predictors were fed into the model, and 15 of them have non-zero relative influence on the smoothed COVID-19 cases. RMSE was used to select the optimal model using the smallest value. The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10. The predictors are ranked according to their relative influence, and the results are presented in Table 3. The covariates with zero influence on the smoothed COVID-19 cases are not considered for further analysis.

Table 3. Ranking of variables based on their relative influence on the smoothed COVID-19 cases.

Variable	Relative influence
NCS	61.47979083
NTS	29.63352677
PR	3.71655489
TC	2.45586959
RR	0.93971713
NT	0.54517464
NDS	0.38453197
NCPM	0.26275310
TT	0.22381632
TV	0.20544590
SI	0.06160743
NV	0.03290140
ND	0.02715463
PFV	0.01611845
NVS	0.01503695
TD	0.00000000
TCPM	0.00000000
TDPM	0.00000000
PV	0.00000000
TB	0.00000000
TVPH	0.00000000

Table 3 shows that the number of new cases per million is highly influential in predicting the number of COVID-19 cases in Namibia. Among the influential factors are the productivity rate (PR), reproduction rate (RR), new vaccines (NV), and stringent measures (SI).

Table 4. ANOVA for the Parametric Effects.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
s(NCPM)	1	44635381	44635381	33467.4669	$< 2.2e-16$ ***
s(NCS)	1	16536599	16536599	12399.0897	$< 2.2e-16$***
s(NTS)	1	58722	58722	44.0295	6.608e-11 ***
s(NT)	1	1408	1408	1.0554	0.3046215
s(PFV)	1	16614	16614	12.4570	0.0004445 ***
s(NVS)	1	7666	7666	5.7482	0.0167746 *
s(RR)	1	9	9	0.0070	0.9331258
s(TT)	1	19477	19477	14.6036	0.0001448 ***
s(TPC)	1	315	315	0.2363	0.6270603
s(PR)	1	62482	62482	46.8489	1.714e-11 ***
s(SI)	1	60048	60048	45.0241	4.103e-11 ***
s(TV)	1	19108	19108	14.3272	0.0001672 ***
s(NV)	1	13875	13875	10.4035	0.0013180 **
s(TC)	1	2723	2723	2.0415	0.1535192
s(NDS)	1	38	38	0.0288	0.8653953
s(ND)	1	21	21	0.0159	0.8998468
Residuals	680	906912	1334		

3.2.3. GAM Using variables selected by the SGBM: The 15 variables selected by the SGBM model are used to fit the generalised additive model (GAM). The GAM is used to analyse the nonlinear effect of each predictor variable. An exponential distribution from a Gaussian family and an identity link function are used to fit the model. Analysis of variance for the parametric effect of the predictors on daily COVID-19 cases is presented in Table 4.

Table 4 shows that the number of people fully vaccinated (PFV), positivity rate (PR), stringent index (SI), total vaccinated (TV), and new deaths smoothed (NCS), among others, have significant nonlinear effects on the daily COVID-19 cases. Although the results are not presented, all variables have a significant nonparametric effect. In Figure 6, the performance of the fitted model on the training dataset is presented. The fitted model managed to capture all the ups and downs in the training dataset.

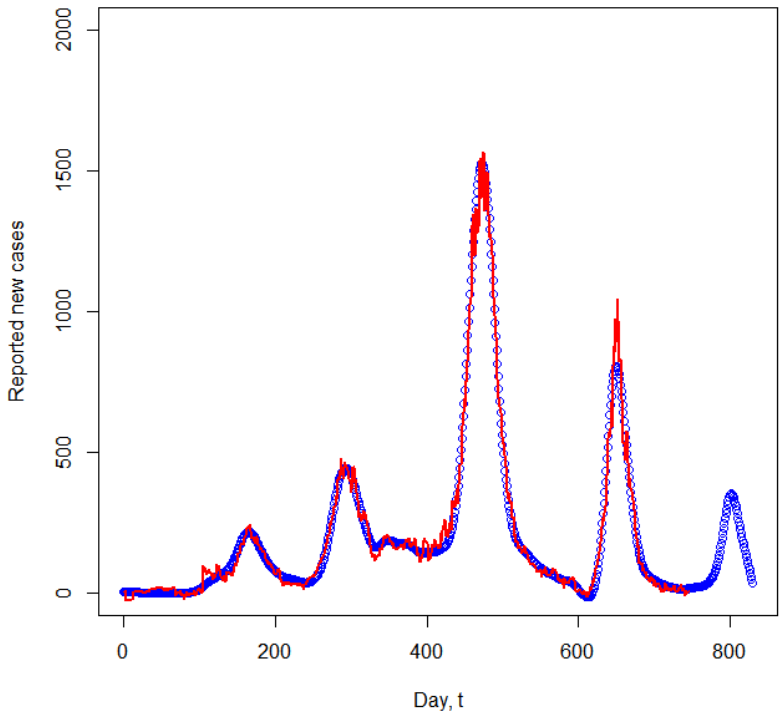

Figure 6. Training Prediction from a GAM model.

3.2.4. SVR model using variables selected by the SGBM approach: The SVMs are fitted for the training data sets using variables selected by the SGBM approach. The four Kernels are considered, that is, the linear, polynomial, radial, and sigmoid kernels. The SVM type used for all the functions is the eps-function. The parameters for the fitted models were gamma=0.0909, cost parameter =1, and epsilon=0.1. For the linear, radial, polynomial, and sigmoid functions, the number of support vectors is 8, 96, 297, and 728, respectively. The polynomial fitted is of degree 3.

In Figure 7, the plots of the fitted SVM models (linear: top left, radial: top right, polynomial: bottom left, and sigmoid: bottom right) on the training dataset are presented. The linear function has RMSE=59.73, MAE=34.71, and 96% of the variability in the daily smoothed COVID-19 cases is explained by the independent variables. The radial function has RMSE=43.9, MAE=23.25, and 98% of the variability in the daily smoothed COVID-19 cases is explained by the independent variables in the model. The polynomial function has RMSE=54.19, MAE=33.57, and 97% of the variability is explained by the independent variables. The performance of the radial kernel function on the training set outperforms all the other kernel functions. The sigmoid kernel function produced the worst performance as shown by very large RMSE and MAE values in Figure 7.

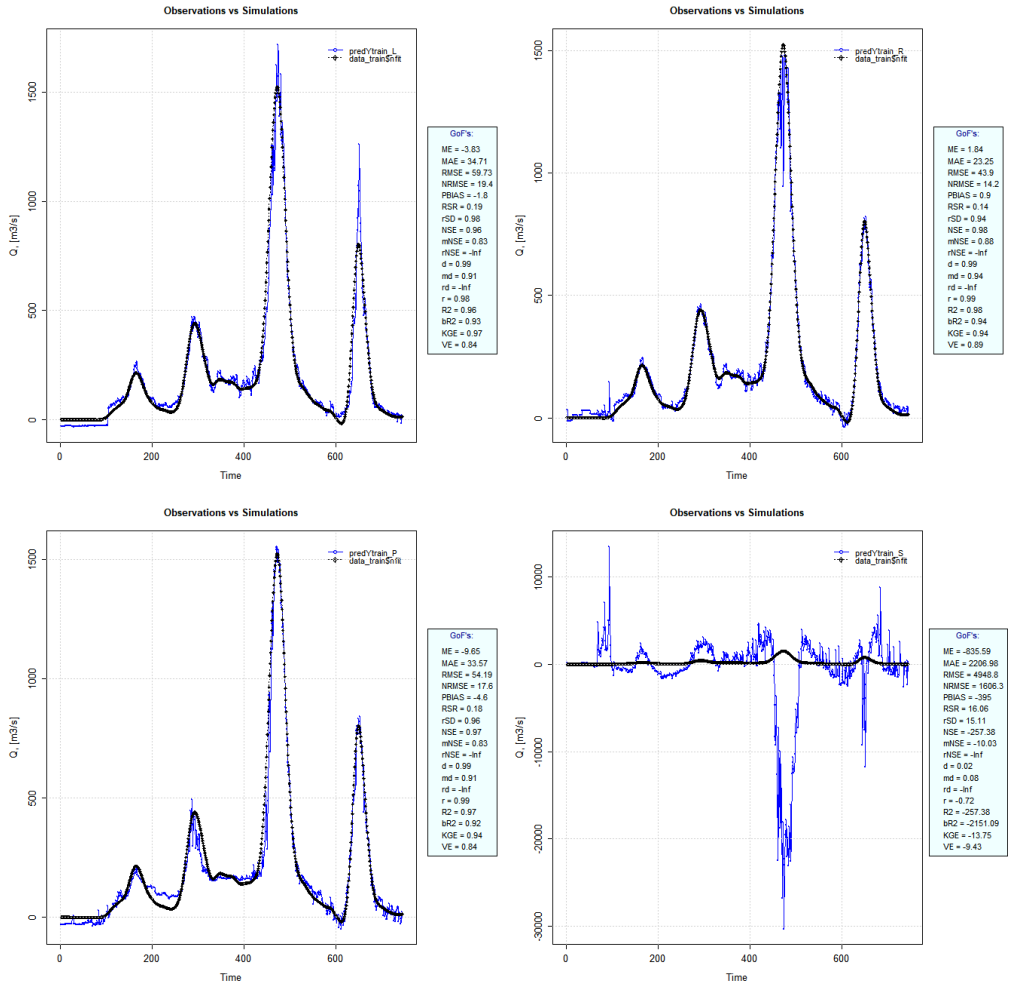

Figure 7. Plots of the fitted models on the training dataset.

3.3. Comparison of the fitted models at different rolling windows

In this section, we validate the fitted models on different rolling windows, that is, 84-, 60-, 30-, and 14-day rolling windows. The error metrics, RMSE, ME, MAE and MAPE, are used to assess the robustness of each fitted model. However, for the comparison of the overall performance of the fitted models, the MAE is used since it is the best error metric for nonlinear data. Nonlinearity models sometimes lead to large, infrequent errors. MAE provides a more stable average in such cases. An alternative is the MAPE.

Table 5 shows that at the 84-day rolling window, the SVM model based on the radial Kernel function outperforms all the other models. The second best is the SGBM, followed by the GAM. At the 60-day rolling window, the best performing model is the TBATS model, followed by the SGBM, and on position 3 is the SVM based on the linear Kernel function. The GAM is superior to all the models in the 30-day rolling window, followed by the SVM linear kernel function. For the 14-day rolling window, the TBATS has proved to be the most superior compared to all the other models, followed by the SVM linear Kernel and then the GAM. The results presented in Table 5 show that machine learning approaches are superior for long-term forecasting. The TBATS model had the worst performance in the long-term horizon. However, for short-term forecasting, the TBATS model is the best model to use. A visual appeal of the performance of the best 3 models from each rolling window is presented in Figure 8.

Table 5. Rolling Window Model Error Metrics

Model	Metric	Rolling Window			
		84 (last wave)	60 (2 months)	30 (1 month)	14 (2 weeks)
TBATS	RMSE	188.1455	85.69587	169.1345	15.67613
	MAE	231.6585	**65.16128**	**118.7662**	**10.53495**
	MAPE		39.80765	157.2638	20.4867
SGBM	RMSE	71.29868	111.7078	182.8663	121.4935
	MAE	**103.0539**	**88.36401**	150.9392	113.3241
	MAPE		51.809	99.8852	138.24
GAM	RMSE	77.90493	125.9679	124.1408	47.98212
	MAE	**106.1451**	104.3015	**95.26497**	**40.47291**
	MAPE		66.39143	46.6215	52.2991
SVM-Linear	RMSE	72.53	125.68	162.71	41.34
	MAE	107.52	**93.51**	**117.11**	**33.09**
	ME	-14.32	-25.00	34.71	-15.09
SVM-Radial	RMSE	81.36	110.07	177.04	106.74
	MAE	**98.37**	97.35	143.18	86.32
	ME	33.02	-8.09	117.86	84.31
SVM-Poly	RMSE	101.23	195.81	222.6	80.89
	MAE	175.4	108.26	151.49	64.09
	ME	9.4	-40.81	75.74	50.99
SVM-Sigmoid	RMSE	997.54	2265.68	2781.15	
	MAE	1850.08	1216.3	1577.38	
	ME	-413.99	-867.02	-1086.00	

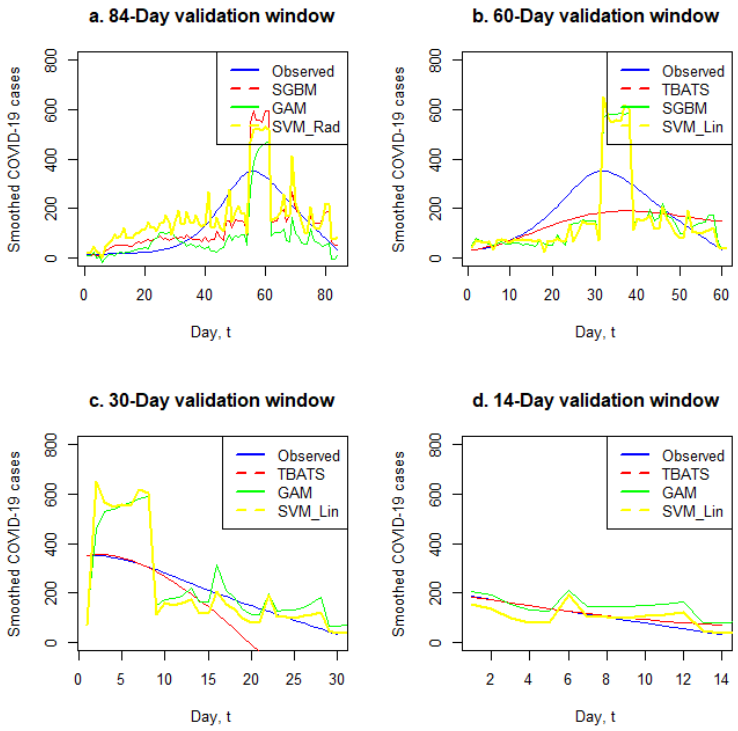

Figure 8. Validation of 3 best fitting models from each of the 84-, 60-, 30-, and 14-day rolling windows.

3.3.1. The SVR(Radial)-GAM-SGBM Combined forecast model for the 84-day forecasting horizon The top three performing models in the 84-day forecasting horizon were the SVM-radial, GAM, and the SGBM models. We use linear quantile regression averaging (LQRA) at different quantiles ($\tau = 0.5, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95$) to combine forecasts from the three best-performing models, thus coming up with a robust SVR(Radial)-GAM-SGBM model. Increasing the quantiles reduced the models' performance. Hence, we present results from the two best-performing quantiles, namely, $\tau = 0.50, 0.55$. Results from the fitted SVR(Radial)-GAM-SGBM model are presented in Table 6 below. For $\tau = 0.50$, forecasts for the radial kernel function do not contribute significantly to the fitted combined forecast model at the 10% level of significance (p-value=0.12453). At the 10% level of significance, forecasts from all the combined models contributed significantly (p-values below 0.1).

Table 6. Contribution of each of the forecasts to the fitted SVR(Radial)-GAM-SGBM Combined forecast model at different quantiles.

	Value	Std. Error	t-value	Pr($>\|t\|$)
$\tau = 0.50$				
(Intercept)	1.16071	10.94162	0.10608	0.91578
sgbm.forecast	1.11482	0.32049	3.47847	0.00082
gam.forecast	-1.04765	0.46918	-2.23295	0.02835
forecastsvm-R	0.30485	0.19638	1.55233	0.12453
$\tau = 0.55$				
(Intercept)	0.89340	12.39582	0.07207	0.94272
sgbm.forecast	1.33731	0.37497	3.56648	0.00061
gam.forecast	-1.45626	0.44214	-3.29366	0.00147
forecastsvm-R	0.41187	0.23467	1.75510	0.08307

The performance of the SVR(Radial)-GAM-SGBM models on the test dataset are presented in Table 7. The results show that forecasts from the 50^{th} quantile outperform forecasts from the 55^{th} quantile based on the MAE, MAPE and Theil's U statistic. Compared to the single-forecast models (SVM-Radial, SGBM, and GAM), the SVR(Radial)-GAM-SGBM model demonstrated its superiority in forecasting daily COVID-19 cases for Namibia.

Table 7. Key performance indicators for the SVR(Radial)-GAM-SGBM Combined forecast model at different quantiles.

Fitted model	ME	RMSE	MAE	MAPE	Theil's U
$\tau = 0.50$					
Test set	19.31754	78.96442	54.50386	59.84571	10.39652
$\tau = 0.55$					
Test set	6.282535	78.52592	54.93769	71.667	13.13742

Plots of forecasts from the SVR(Radial)-GAM-SGBM model superimposed on the observed test set from the 50^{th} and the 55^{th} quantiles are presented in Figure 9. A comparison of these plots show that the SVR(Radial)-GAM-SGBM model performs better at the 50^{th} quantile. This conforms to the findings from Table 7.

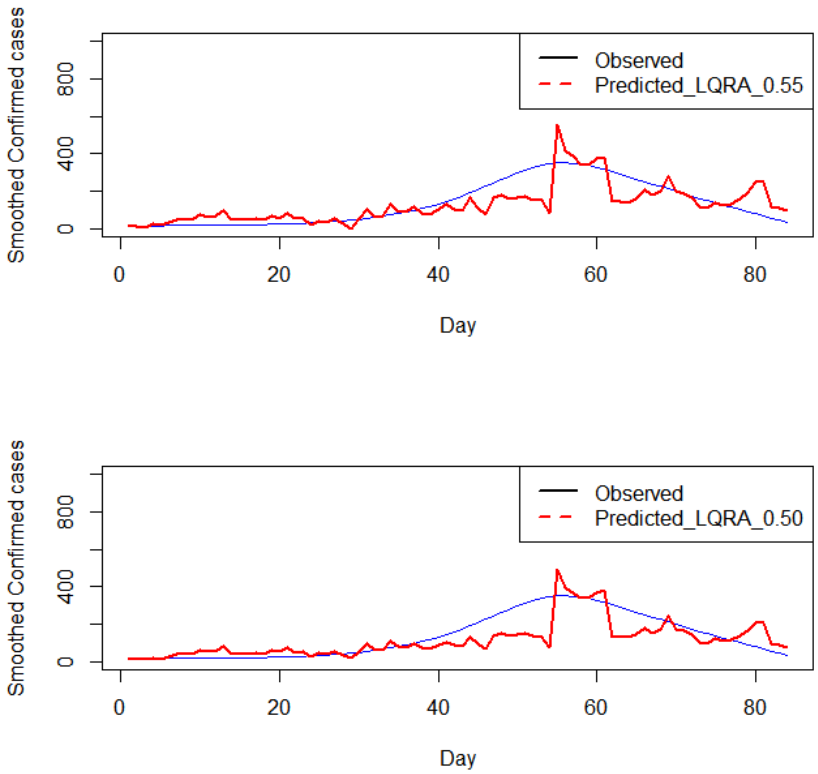

Figure 9. Plots for the performance of forecasts from the combined on the test dataset from the smoothed time series at the 55th and 50th quantiles.

4. Discussion

This research focused on using machine learning algorithms to predict the future dynamics of COVID-19 in Namibia. The time series data for the daily COVID-19 cases are nonlinear and not normally distributed. Normalisation of the data was done by fitting a cubic spline model to the data and then extracting the predicted values, which resulted in a smoothed time series with smoothing parameter $\lambda = 0.472716$. Previous time series [26]; [27] proposed a log-transformation approach to normalise positively skewed data, thereby making the distribution more closely resemble a normal distribution, a common requirement for statistical analysis. However, with the COVID-19 data, the log-transformation could not work because on some days the reported number of cases was zero and the logarithm of zero is undefined, leading to the distortion of results. The main machine learning learning models applied on the smoothed time series were the stochastic gradient boosting machine (SGBM) model to identify the influential variables, the generalised additive model (GAM), and the support vector machine (SVM) model using the kernel transformations (linear, radial, polynomial, and sigmoid). These models were trained on 90% of the time series data and validated on the remaining 10% whilst benchmarking their performance on the TBATS model. The 10% validation set represents the last wave of COVID-19 in Namibia. In addition, the fitted models were also validated on a rolling window that included 60- (2 months), 30- (1 month), and 14- (2 week) day-ahead. Compared to previous studies [28], this study incorporated external factors that include vaccination rates, stringent index, reproduction rate, etc. These factors helped to improve the prediction accuracy of the fitted models and provide a more complete view of the COVID-19 dynamics. The choice of the machine learning approaches,

the SVR, GAM, and SGBM, which help to solve regression time series problems, was motivated by the need to incorporate some external factors in the models. The best three forecast models on the last wave (10%) were combined using the LQRA approach, and this resulted in a robust SVR(radial)-GAM-SGBM combined forecast model with optimum performance at the 50^{th} quantile, which is similar to the median.

The TBATS model showed that the time series for the daily COVID-19 cases does not have seasonal variations.. [29] argued that this was because the spread was mainly impacted by human behaviour and public health interventions that distorted the natural transmission patterns. The interventions include vaccinations, lockdowns or stringent measures. Kraamwinkel [30] argued that in a pandemic case, data does not clearly show a clear trend, making it hard to extract data patterns. Thus, during the periods of higher stringent indices, resulted in lower incidence of COVID-19. Also, the vaccinated individuals become immune to COVID-19 infection, thereby reducing the number of new infections. Analysis from the SGBM picked stringent index and vaccine-related variables among the variables that had a relative influence on the daily COVID-19 cases. Results from the GAM also showed that stringent index and vaccine-related variables had significant parametric effects on the spread of the disease.

From the rolling windows that were used, the SVM radial kernel outperformed all the fitted single models in the 84-day forecasting horizon, agreeing with findings from previous findings [31]. A study that was carried out using data from Zimbabwe also supports the use of the SVR approach, although in the study, the radial kernel function had the best performance [25]. For shorter horizons, especially the 14-day forecasting, the TBATS model demonstrated its superiority over machine learning models. This finding supports results from the literature that traditional statistical models struggle to capture complex, nonlinear relationships and fail to account for external factors that become more influential over longer horizons. The TBATS model relies on historical patterns, thus assuming that the patterns continue predictably into the future. [32] argued that sigmoid functions perform the best when the number of classes becomes very large, in which the linear function becomes less precise. Thus, contradicting the findings from this study. In predicting the spread of COVID-19 in Zimbabwe, [25] also observed that the radial kernel function was the best. However, unlike this study, the study by Shoko and Sigauke only focused on one validation set, which was prediction in the long run. This study further demonstrated the power of combined forecasting based on the probabilistic approach for long-term forecasting. This approach has proved that integrating forecasts from different models improves the predictive power, particularly at the 50^{th} quantile. A combined forecast approach leverages the strengths of multiple models, resulting in reduced errors. Therefore, this study demonstrates that smoothed time series give meaningful results in cases where the data is non-normal and has some zero values. The study further recommends the use of combined forecasts, in particular the SVR(radial)-GAM-SGBM ensemble algorithm, to forecast epidemics that have similar dynamics as COVID-19. Machine learning approaches have, overall, proven their superiority over traditional approaches, such as the TBATS model. Findings from these results can be used by the health sector for the proper management and control of COVID-19 epidemiology, as well as to create awareness and help as an early warning sign for the pandemic.

5. Conclusion

This study proposes the inclusion of external factors to improve the forecasting performance of models. Although the prediction of the future with accurate forecasting remains difficult to impossible, this study has demonstrated the potential of machine learning models for long-term forecasting and the classical TBATS model for short-term forecasting. For a more robust performance for long-term forecasting, combining machine learning forecast models is advisable. This paper addresses a critical need for accurate COVID-19 forecasting in understudied regions like Namibia and supports SDG 3 of health and wellness.

5.1. Limitation

The statistical analysis was based on secondary data, and we did not have control over the original data collection methodology. More informative results could have been reached if primary data had been used.

Acknowledgment

The authors would like to acknowledge Our World in Data for making the data openly available for use. We would also like to extend our acknowledgement to the reviewers for their contributions, which helped improve the quality of our paper.

REFERENCES

1. Samuel Adu-Gyamfi, Edward Brenya, Razak M. Gyasi, Kabila Abass, Benjamin Dompreh Darkwa, Michael Nimoh and Lucky Tomdi. *A COVID in the wheels of the world: A contemporary history of a pandemic in Africa.* Research in Globalisation, 2021.
2. Farrukh Saleem, Abdullah Saad AL-Malaise AL-Ghamdi, Madini O. Alassafi and Saad Abdulla AlGhamdi , *Machine Learning, Deep Learning, and Mathematical Models to Analyse Forecasting and Epidemiology of COVID-19: A Systematic Literature Review*, International Journal of Environmental Research in Public Health (2022), doi:https://doi.org/10.3390/ijerph19095099
3. Mateus Maia, Jonatha S. Pimentel, Ivalbert S. Pereira, João Gondim, Marcos E. Barreto and Anderson Ara, *Convolutional Support Vector Models: Prediction of Coronavirus Disease Using Chest X-rays*, Information. doi:10.3390/info11120548
4. World Health Organisation *Coronavirus Disease (COVID-19) Dashboard*, https://covid19.who.int/ (accessed on 23 February 2023)
5. Hugo Capell a Miternique, *SARS-CoV-2 effects on tourism. The recovery of regional complexity: When less means more: The case of Balearic Islands* , Research in Globalisation (2021).
6. Ashutosh Trivedi, Nanda Kishore Sreenivas, and Shrisha Rao, *Modeling the Spread and Control of COVID-19*, Systems, 2021. https://doi.org/10.3390/systems9030053
7. Showmick Guha Paul, Arpa Saha, Al Amin Biswas , Md. Sabab Zulfiker, Mohammad Shamsul Arefin, Md. Mahfujur Rahman , Ahmed Wasif Reza, *Combating Covid-19 using machine learning and deep learning: Applications, challenges, and future perspectives*, Array 2023
8. Ndeyapo M. Nickanor, Godfrey Tawodzera and Lawrence N. Kazembe, *The Threat of COVID-19 on Food Security: A Modelling Perspective of Scenarios in the Informal Settlements in Windhoek*, Land (2023), doi:https://doi.org/10.3390/land12030718
9. Ozili, P.K. *COVID-19 in Africa: Socio-Economic Impact, Policy Response, and Opportunities*, Int. J. Sociol. Soc. Policy 2020, 42, 177–200. Available online: https://www.emerald.com/insight/0144-333X.htm (accessed on 4 August 2020).
10. UN-HABITAT, *COVID-19: Socio-Economic Impacts in Africa*, Discussion Paper-9 April 2020. 2020. Available online: https://unhabitat.org/sites/default/files/2020/04/dp_covid-19_effects_in_africa5.pdf (accessed on 4 August 2020)
11. Godfrey Bwire, Alex Riolexus Ario , Patricia Eyu, Felix Ocom, Joseph F. Wamala4, Kwadwo A. Kusi , Latif Ndeketa6 , Kondwani C. Jambo6,7 , Rhoda K. Wanyenze2 and Ambrose O. Talisuna, *The COVID-19 pandemic in the African continent*, BMC Medicine, 2022. https://doi.org/10.1186/s12916-022-02367-4
12. Kossi Amouzouvi , Kétévi A. Assamagan, Somiéalo Azote, Simon H. Connell , Jean Baptiste Fankam Fankam, Fenosoa Fanomezana , Aluwani Guga, Cyrille E. Haliya, Toivo S. Mabote , Francisco Fenias Macucule , Dephney Mathebula, Azwinndini Muronga, Kondwani C.C. Mwale , Ann Njeri , Ebode F. Onyie, Laza Rakotondravohitra, George Zimba, *A model of COVID-19 pandemic evolution in African countries*, Scientific African, 2021. https://doi.org/10.1016/j.sciaf.2021.e00987
13. Jaspreet Kaur, and Prabhpreet Kaur. *COVID-19 Future Forecasting based on Time-series statistical analysis using Machine Learning Model*, International Conference on Smart Systems and Advanced Computing (Syscom-2021), December25–26, 2021.
14. Daniel Andrade-Girón, Edgardo Carreño-Cisneros, Cecilia Mejía-Dominguez, Julia Velásquez-Gamarra, William Marín-Rodriguez, Henry Villarreal-Torres, Rosana Meleán-Romero *Support vector machine with optimized parameters for the classification of patients with COVID-19*, EAI Endorsed Transactions on Pervasive Health and Technology, 2023.
15. Iqbal H. Sarke , *Machine Learning: Algorithms, Real-World Applications and Research*, Directions (2021). Springer Nature Computer Science. https://doi.org/10.1007/s42979-021-00592-x
16. Castillo, O., Castro, J.R. & Melin, P. *Forecasting the COVID-19 with Interval Type-3 Fuzzy Logic and the Fractal Dimension.* Int. J. Fuzzy Syst. 25, 182–197 (2023). https://doi.org/10.1007/s40815-022-01351-7
17. Yogesh Kumar, Apeksha Koul, Sukhpreet Kaur, Yu-Chen Hu. *Machine Learning and Deep Learning Based Time Series Prediction and Forecasting of Ten Nations' COVID-19 Pandemic*, SN Computer Science, 2022. https://doi.org/10.1007/s42979-022-01493-3
18. Yassine Meraihi, Asma Benmessaoud Gabis, Seyedali Mirjalili, Amar Ramdane-Cherif5, Fawaz E. Alsaadi, *Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey*, SN Computer Science. https://doi.org/10.1007/s42979-022-01184-z
19. Laureen Mueeddin *Real-Time Assessment (RTA) of UNICEF's Ongoing Response to COVID-19 in Eastern and Southern Africa.*, Oxford Management Policy (2021).
20. Namibia Statistics Agency, *A year living with COVID-19: COVID 19 Households and Job Tracker Survey (2021)*, (2021)
21. Tibshirani R, *Regression shrinkage and selection via lasso. Journal of the Royal Statistical Society*, Series B (methodology), 1996, 58(1), 267-288. https://statweb.stanford.edu/~tibs/ftp/lasso-retro.pd
22. Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K. and Simon, N. Lasso and Elastic *Net Regularized Generalized Linear Models: glmnet r package version 4.1-2,*, 2021. https://cran.r-project.org/web/packages/glmnet/glmnet.pdf (Accessed on 10 February 2023)

23. Jalila Jbilou and Salaheddine El Adlouni, *Generalized Additive Models in Environmental Health: A Literature Review*, Novel Approaches and Their Applications in Risk Assessment. April 2012. DOI: 10.5772/38811. https://www.researchgate.net/publication/224830449.
24. Aruna S and Rajagopalan SP, *A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer*, nt J Comput Appl 31(8): 14-20, 2011.
25. Claris Shoko, Caston Sigauke, *Short-term forecasting of COVID-19 using support vector regression: An application using Zimbabwean data*, American Journal of Infection Control (2023), https://doi.org/10.1016/j.ajic.2023.03.010.
26. Leigh Metcalf, William Casey, *Chapter 4 - Introduction to data analysis,*, Cybersecurity and Applied Mathematics, Syngress, 2016, Pages 43-65, https://doi.org/10.1016/B978-0-12-804452-0.00004-X.
27. FENG, C., WANG, H., LU, N., CHEN, T., HE, H., LU, Y., & TU, X. M. *Log-transformation and its implications for data analysis*, Shanghai Archives of Psychiatry, 26(2), 105. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009
28. Yang, W., & Chang, X., *Time series analysis and prediction of the trends of COVID-19 epidemic in Singapore based on machine learning*, Computer Methods and Programs in Biomedicine Update, 7, 100190. https://doi.org/10.1016/j.cmpbup.2025.100190
29. Esposito, M. M., Turku, S., Lehrfield, L., & Shoman, A., *The Impact of Human Activities on Zoonotic Infection Transmissions. Animals* , An Open Access Journal From MDPI, 13(10), 1646, 2023. https://doi.org/10.3390/ani13101646
30. Kraamwinkel S *Time Series Forecasting on COVID-19 Data and Its Relevance to International Health Security [Internet]*, Contemporary Developments and Perspectives in International Health Security - Volume 3. IntechOpen; 2022. Available from: http://dx.doi.org/10.5772/intechopen.104920
31. Shujun Huang, Nianguang, Guang Cai, Pedro Penzuti Pacheco, Shavira Narandes, Yang Wang, and Wayne Xu *Applications of Support Vector Machine (SVM) Learning in Cancer Genomics*, Cancer Genomics and Proteomics 15: 41-51 (2018). doi: 10.21873/cgp.20063
32. enajiba Yassin, Chrayah Mohamed, Al-Amrani Yassine, *A Nonlinear Support Vector Machine Analysis Using Kernel Functions for Nature and Medicine*, E3S Web of Conferences 319, 01 (2021). https://doi.org/10.1051/e3sconf/202131901103.
33. B.E.Boser, I.M.Guyon, and V.N.Vapnik, *A training algotighm for optimal margin classifiers.*, Proceedings of the 5th Annual ACM workshop on Computational Learning, pages 144–152, 1992.