Journal of Applied Mathematics and Physics, 2024, 12, 2904-2917
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

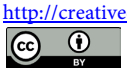
Predicting Malaria Dynamics in Burundi Using Deep Learning Models

Daxelle Sakubu[1*], Kelly Joelle Gatore Sinigirira[1,2], David Niyukuri[1,2,3]

[1]Doctoral School, University of Burundi, Bujumbura, Burundi
[2]Department of Mathematics, University of Burundi, Bujumbura, Burundi
[3]The South African Department of Science and Technology-National Research Foundation, (DST-NRF) Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Cape Town, South Africa
Email: *daxelle.sakubu@ub.edu.bi, kelly.gatore@ub.edu.bi, david.niyukuri@ub.edu.bi

How to cite this paper: Sakubu, D., Gatore Sinigirira, K.J. and Niyukuri, D. (2024) Predicting Malaria Dynamics in Burundi Using Deep Learning Models. *Journal of Applied Mathematics and Physics*, **12**, 2904-2917.
https://doi.org/10.4236/jamp.2024.128173

Received: June 19, 2024
Accepted: August 19, 2024
Published: August 22, 2024



Abstract

Malaria continues to be a major public health problem on the African continent, particularly in Sub-Saharan Africa despite the ongoing efforts and significant progress that has been made. In the case of Burundi, malaria remains a major public health concern in the general population. In the literature, there are limited malaria prediction models for Burundi knowing that such tools are much needed for intervention design. In this study, deep-learning models are built to estimate malaria cases in Burundi. The forecast of malaria cases was carried out both at the provincial and national levels. Long short term memory (LSTM) model, a type of deep learning model, has been used to achieve best results using climate-change related factors such as temperature, rainfall, relative humidity, together with malaria historical data and human population. With this model, the results showed that different parameter tuning can be used to determine the minimum and maximum expected malaria cases. The univariate version of that model (LSTM), which learns from previous dynamics of malaria cases, gives more precise estimates, but both univariate and multivariate models have the same overall trends at the province level and country level.

Keywords

Malaria, Prediction, Deep Learning, Long-Short-Term Memory (LSTM), Burundi

1. Introduction

Malaria is an infectious disease caused by the Plasmodium falciparum parasite, transmitted by the bite of the female Anopheles mosquito. There are four main

types of malaria parasites: Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, and *Plasmodium ovale*. Plasmodium falciparum is the most dangerous, with a long incubation period of 6 - 14 days [1], and it is the primary cause of malaria in Burundi, accounting for nearly 90% of cases [2].

According to the World Health Organisation (WHO) report [3], estimated malaria cases increased from 213 million to 228 million, and deaths increased from 534,000 to 602,000 in the WHO African Region between 2019 and 2020. This region accounted for approximately 95% of all cases and 96% of all deaths globally, children under the age of five account for 80% of all deaths in this region.

In Burundi, malaria infection is among the main public health concerns after diarrhoea diseases, neonatal disorders, and tuberculosis [3]-[6]. During the recent malaria outbreak in 2017 [7]-[9], World Vision International assumed that climate change, population density, change in agricultural methods, food shortages and a lack of information and action to prevent malaria were the main driving factors contributing to the prevalence of the disease in Burundi [10].

On a global scale, coordinated initiatives to minimize the malaria epidemic are being planned as part of the millennium development goals. In Burundi, the package of interventions for malaria control during pregnancy includes the promotion and use of ITN, IPTp with sulfadoxine-pyrimethamine under directly observed treatment, and fast and successful treatment. The NMCP has yet to follow WHO recommendations from 2016, which increased the recommended number of prenatal care visits from four to eight. With Malaria Operational Plan (MOP) funds from Fiscal Year (FY) 2021 [11], the team proposes to trial an evidence-based group prenatal care strategy to enhance IPTp uptake (according to DHIS2, 54% of women received the recommended three doses at U.S. Government-supported health centers in the first quarter of FY 2021). Although massive efforts have already been deployed, Malaria still prevails. Thus, this project aims to understand the contributions of climate by using machine learning models to predict the cases of Malaria.

In recent years, there has been a great deal of interest in the development and application of machine learning (ML) in the field of infectious diseases [12]. Not only as a catalyst for academic studies but also as a critical means of detecting pathogenic microorganisms, implementing public health surveillance, investigating host-pathogen interactions, discovering drug and vaccine candidates, etc. According to one survey, ML is used in 77% of the products we use today. Machine learning (ML) is a subfield of Artificial Intelligence (AI) that is an important tool in bioinformatics [13]. When confronted with a range of large and complex data sets that must be analyzed, ML may employ sophisticated algorithms and efficient models to extract meaningful information from vast amounts of complex data-sets [14]-[16]. Machine learning extracts useful information from enormous amounts of data by using algorithms to recognise patterns and learn in an iterative process. Instead of relying on any preconceived

equation that may serve as a model, ML algorithms use computing methods to learn directly from data [17]. The union of mathematics and computer science in ML has shown significant potential as a breakthrough in science and technology, and it has been applied to a wide range of scientific fields, including biology. Deep learning models which are part of machine learning relies on multiple hidden layers to learn powerful representations of the input data that has been used in order to predict a DNA sequence function [18], the authors used a convolution neural network combined with LSTM while they used random dropout for improvement. However, the study was limited in its ability to determine the optimal range of dropout.

A number of studies have already been conducted on the prediction of malaria in Burundi, and machine learning techniques, such as artificial neural networks, have been applied [19]. The authors investigated malaria in different groups of ages and the impact of meteorological factors on the high number of malaria cases during some seasons. Different degrees of precision were reported from previous investigations. Nevertheless, the paper did not show the prediction on a province level which may contribute during intervention or emergency cases on what part of the country needs more attention in a specific moment. The overarching goal of this research is to investigate malaria case predictions based on meteorological data in order to help future intervention teams on what part of the country may need more attention due to its high number of malaria cases that may contribute to the country becoming an endemic situation. Hence, LSTM will be used to forecast the incidence of malaria in all five provinces. This study may help improve public health measures, especially at the district level.

2. Material and Methods

In the following section, we discuss the data we used and how we built the deep learning models. All data used was from Burundi, and data analysis, and the models were processed using Python Language (version 3.6.5) [20].

It is worth noting that, before choosing the best deep learning model we used in this work, we explored so many different deep learning models such as decision trees and artificial neural. However, these alternative models were unable to achieve the same level of accuracy as the LSTM model.

2.1. Data Description

The study was carried out using monthly data, collected from different sources, namely: Geographical Institute of Burundi (Institut Geographique du Burundi, IGEBU), the Institute of Statistics and Economic Studies of Burundi Institut de Statistiques et d'Etudes Economiques du Burundi, ISTEEBU and Burundi's National Malaria Control Programme (NMCP). The data was collected for all the eighteen provinces of the country. Since it's national data they are not available for the public.

2.2. Data Extraction

Data collected from IGEBU were on a monthly scale from 2010 to 2022 with parameters such as relative humidity, rainfall, and temperature with their maximum and minimum values. The average was calculated and inserted in the data-set. Historical malaria data was obtained at the NMCP Burundi on a monthly scale from January 2010 to December 2022 for all 18 provinces. The human population feature was available online on ISTEEBU website and the human population was calculated annually.

2.3. Data Processing

The data collected was on a different time scale monthly and annually, thus the human population was considered constant throughout the year. The meteorological data contained some missing values that were filled using an algorithm called miss Forest [21]. The experiments were done on 80% of the data set while the testing was done on the rest of the data set which means 20%. The mean squared error was used for loos function and bacthe size was 12. This algorithm was judged to be suitable since it takes into account the possible relation between variables. The data were normalized before being fed into the neural networks. Recently, the communes, zones, and hills/neighbourhoods of the Republic of Burundi have been the subject of administrative regrouping [22]: Bujumbura (west part), Buhumuza (east part), Gitega (central part), Burunga (southern part) and Butanyerera (northern part) are the five provinces of the country in the new delimitation of provinces. This grouping is mainly based on the so-called natural regions of Burundi, which exhibit differences in terms of climate conditions, agriculture, landscape, and social life in those areas.

In this reform, the provinces were regrouped as follow: Bujumbura included all the commune of Bujumbura Mairie, Bujumbura Rural, Bubanza and Cibitoke. Gitega assembled all the commune of Gitega, Mwaro, Karuzi and Muramvya. Buhumuza grouped together all the communes of Cankuzo, Muyinga and Ruyigi. Butanyerera aggregated all the communes of Kirundo, Ngozi and Kayanza. Finally, Burunga included all the communes of Bururi, Makamba, Rumonge and Rutana. Therefore, the data-set of these new delimited provinces were the mean of the old provinces regrouped per month for the meteorological data and the sum for the human population and malaria cases data.

After processing the data, we start to build the neural network models. The prediction model was built with four layers and several units. Nevertheless, the performance was not good, hence a different approach was taken in order to find good results.

2.4. Building ML Models

Since the data set was on a monthly scale, a different model that took into account the previous information seems to be appropriate for this study. The kind of algorithm that takes into account previous information in a chronological

manner is a recurrent neural network. This kind of neural network, unlike others, has feedback connections. The network connection weights and biases change once per training episode, similar to how physiological changes in synaptic strengths store long-term memories; the network's activation patterns change once per time step, similar to how the brain's electrical firing patterns change moment-to-moment to store short-term memories. Recurrent neural networks are frequently used in multiple domains to predict future events based on previous experience. Long-Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is particularly well-suited for handling sequences of data. It was introduced by Hochreiter and Schmidhuber in 1997 [23] and has since become one of the most popular and widely used architectures for tasks involving sequential data. The model was implemented using Tesnsorflow and Keras. The LSTM model architecture consisted of one layer of LSTM with 5 units, followed by a dense layer.

For instance, Long Short Term Memory (LSTM) model can be used for tasks like connected, unsegmented handwriting recognition, video games, speech recognition, automated translation, healthcare, Speech activity detection and Robotics. In the healthcare system, it has been used for cardiovascular prediction analytics, where it yields the best accuracy among other machine learning models [24]. Since LSTM has been shown to perform well in aiding in rational decision making, it was used in this study to predict malaria cases.

In **Figure 1**, a long short term memory cell is depicted where the square represents the layers, the ellipse is the component wise operation, C_i are hidden state vectors and X_i are input vector to the LSTM unit and h_i is hidden state vector also known as output vector y_i of the LSTM unit.

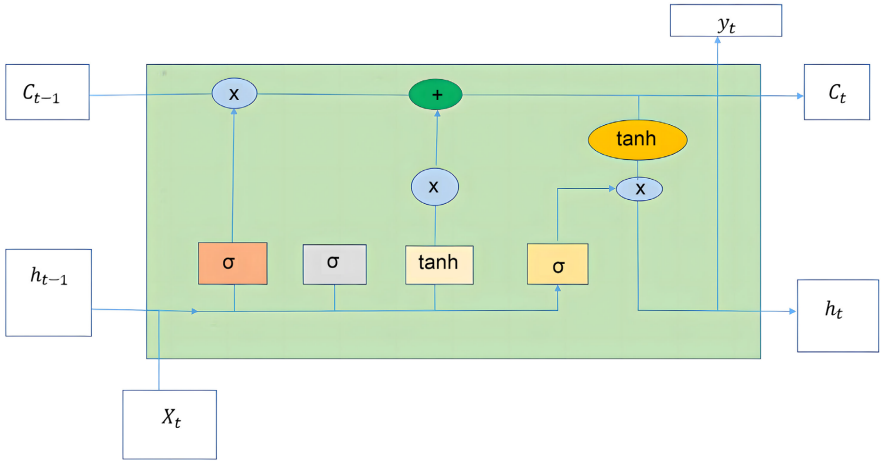


Figure 1. Long short term memory (LSTM) cell.

In **Figure 2**, a univariate LSTM model is shown in a sequential manner with previous states alongside the input vectors x_i and outputs y_i. The previous malaria cases are the input vector and the output vector is the actual malaria cases. The Multivariate LSTM model is represented in **Figure 3** with the multiple

input vector x_i up to x_n and the target y_i. In the multivariate LSTM model the input vectors are the climate data, human population and previous malaria cases, and the output vectors are the current malaria cases.

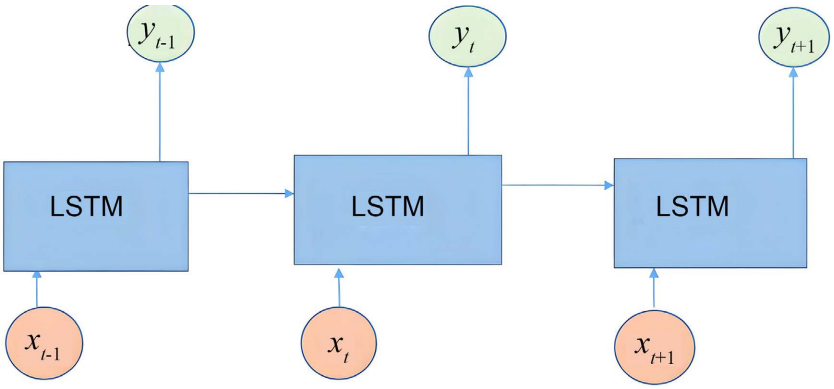


Figure 2. Univariate long short term memory.

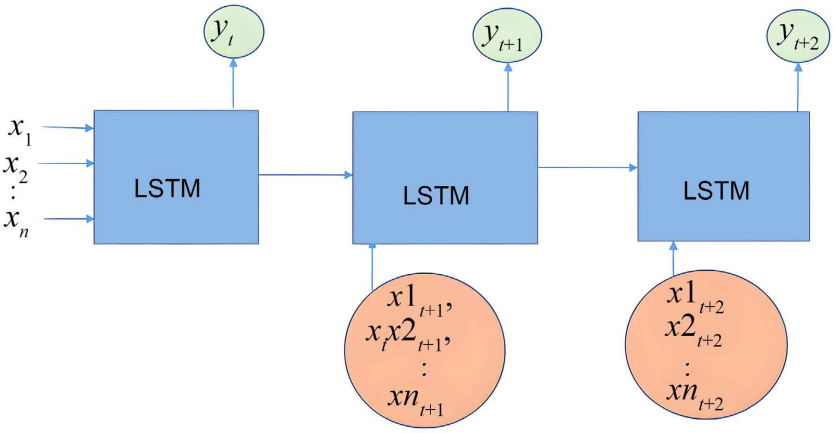


Figure 3. Multivariate long short term memory.

After tuning the data-set in the ML models we obtained results which are presented in the following section.

3. Results

After running and fitting the data to the models, the error between the actual and predicted cases was calculated using the root mean square error (RMSE). From the results obtained are observed in **Table 1**, overall, the univariate LSTM model got the smallest RMSE while the multivariate got the biggest RMSE. In Bujumbura province, the multivariate error is nearly four times as big as the univariate error. In the case of Gitega province, the error gap is three times big between multivariate and univariate and this is the same for Butanyerera province as well. In Burunga province, the multivariate error is twice bigger than the univariate error. For Buhumuza province, the error gap was huge, the multivariate error was about eight times bigger than the univariate one. In the country-level prediction, the multivariate error disparity was three times larger than

the univariate one.

Table 1. RMSE of malaria cases between Univariate LSTM and Multivariate LSTM.

Provinces	RMSE	
	Univariate LSTM	Multivariate LSTM
Bujumbura	4868.69	16777.17
Gitega	10943.18	31012.00
Burunga	6403.33	12964.48
Butanyerera	8288.07	25187.09
Buhumuza	5664.18	44893.25
Country level: Burundi	31635.95	119724.68

Table 2. Number of malaria cases predicted by Univariate LSTM and Multivariate LSTM and observed.

Provinces	Malaria Cases		
	Observed	Univariate LSTM	Multivariate LSTM
Bujubura	2,253,588	2,232,670	2,280,516
Gitega	2,395,666	2,444,657	2,338,582
Burunga	1,901,832	1,890,808	1,974,578
Butanyerera	3,568,702	3,507,692	3,808,420
Buhumuza	2,772,583	2,808,850	3,567,723
Country level: Burundi	12,892,371	12,841,653	15,215,766

3.1. Province-Level Predictions

The prediction of malaria cases on the province level is shown in **Figure 4** from October 2020 to September 2022. In univariate LSTM prediction, the curve trends are followed in most cases with the observed cases being slightly higher than the expected ones. In the multivariate LSTM prediction, the curve trends are not coordinated except for the country level prediction. The difference between the observed and predicted cases is seen in **Table 2**. The total number of malaria cases in Gitega province during that time period was higher than predicted by the multivariate model but lower than predicted by the univariate model. The observed malaria cases in Burunga province were somewhat higher than the multivariate prediction and lower than the univariate prediction. The multivariate model in Buhumuza province had a much greater number of instances than the actual cases, but the observed cases during that time were slightly fewer than the number predicted by the univariate model. The multivariate for Butanyerera was substantially higher than the observed instances, despite the fact that the actual cases were only marginally higher than the univariate forecast. The multivariate LSTM model predictions were marginally higher than

the actual instances, however the observed malaria cases in the Bujumbura province were slightly higher than the Univariate LSTM model projections. (Figures 4-13)

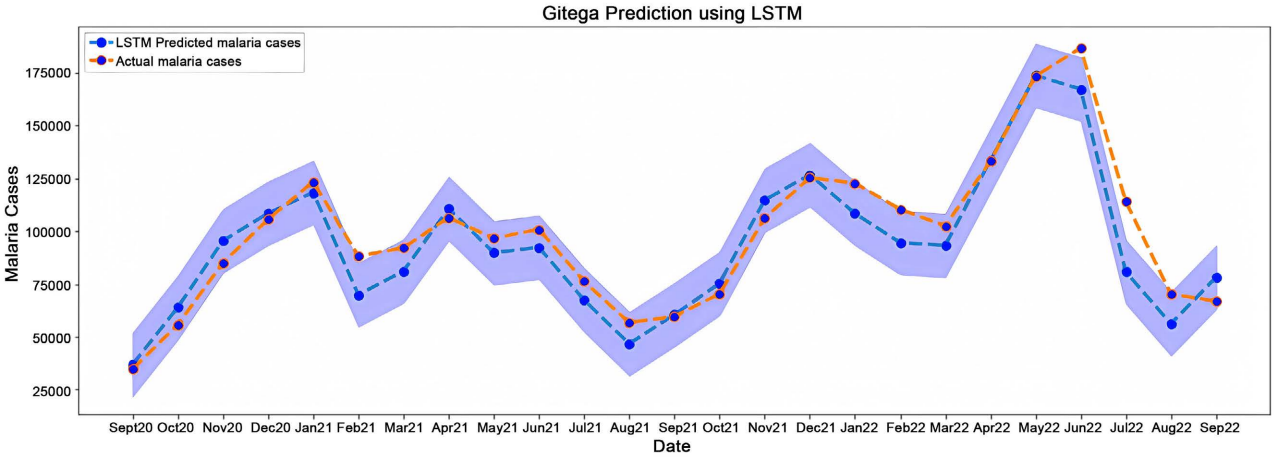


Figure 4. Univariate LSTM Gitega Prediction.

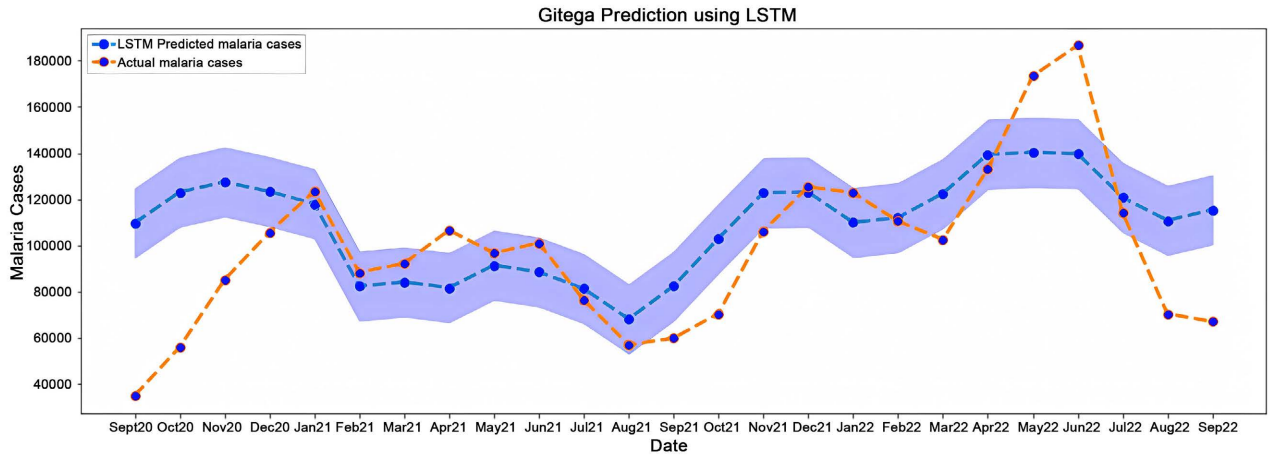


Figure 5. Multivariate LSTM Gitega Prediction.

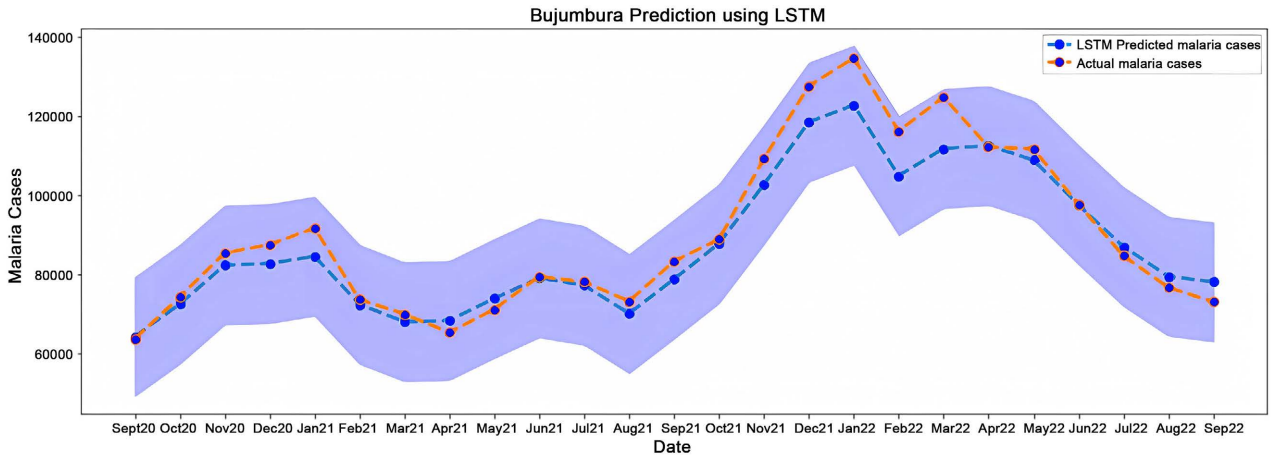


Figure 6. Univariate LSTM Bujumbura Prediction.

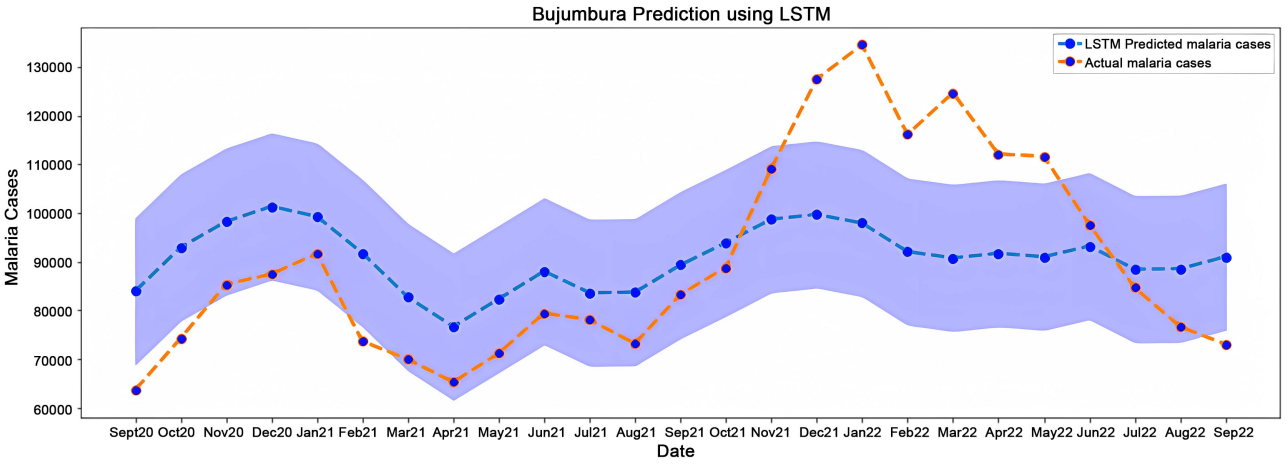


Figure 7. Multivariate LSTM Bujumbura Prediction.

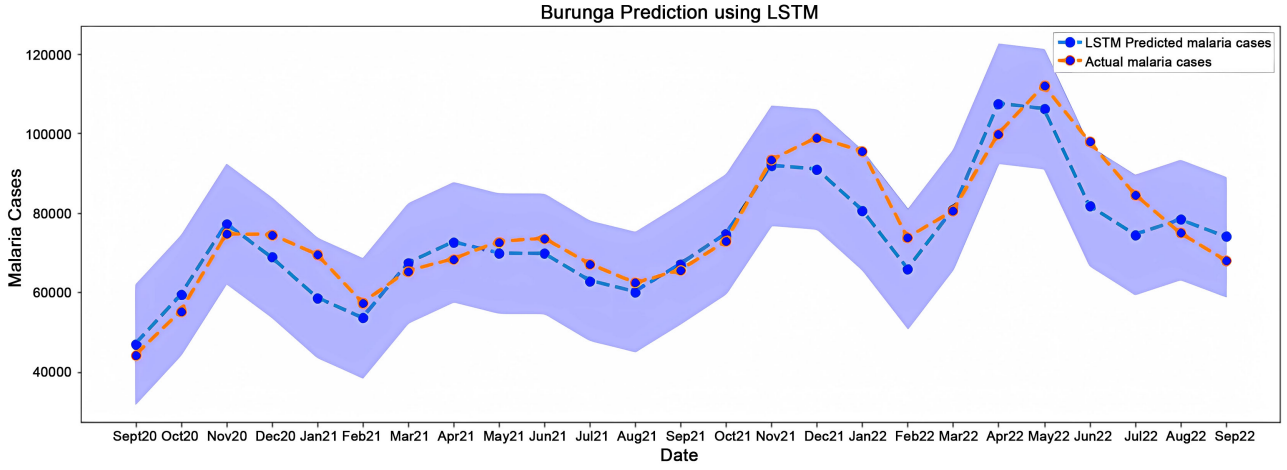


Figure 8. Univariate LSTM Burunga Prediction.

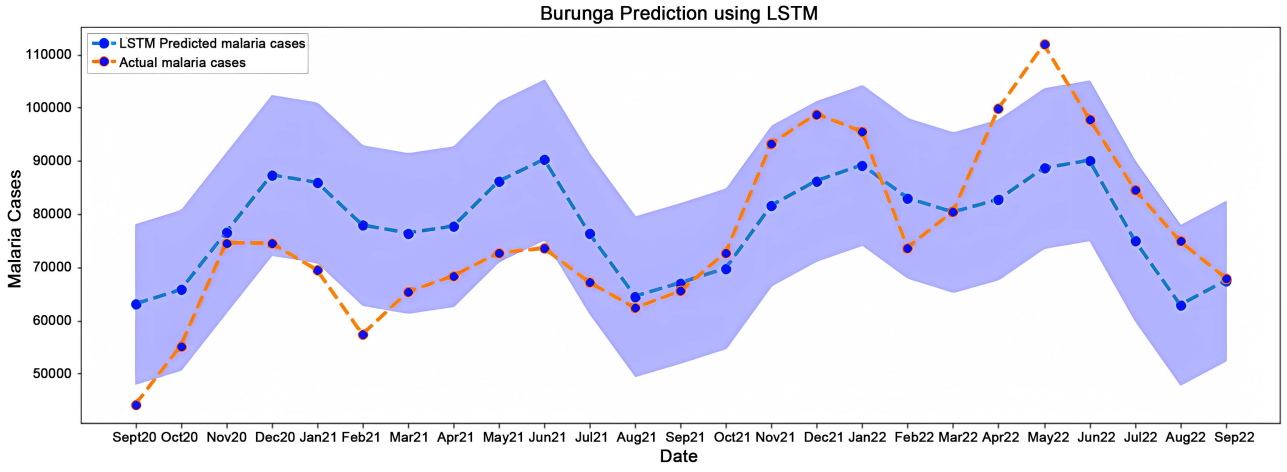


Figure 9. Multivariate LSTM Burunga Prediction.

3.2. Country-Level Predictions

After predicting malaria cases at the provincial level, the country-level predic-

tion was made. The results are depicted in Figure 14 and Figure 15. The data-set of the country level was the mean of all provinces for the climate data and the

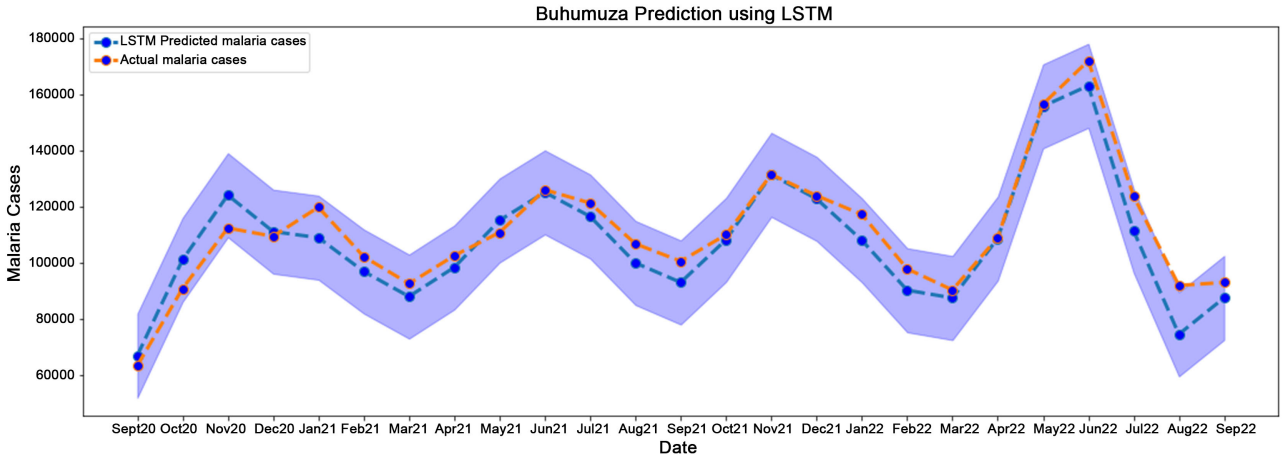


Figure 10. Univariate LSTM Buhumuza Prediction.

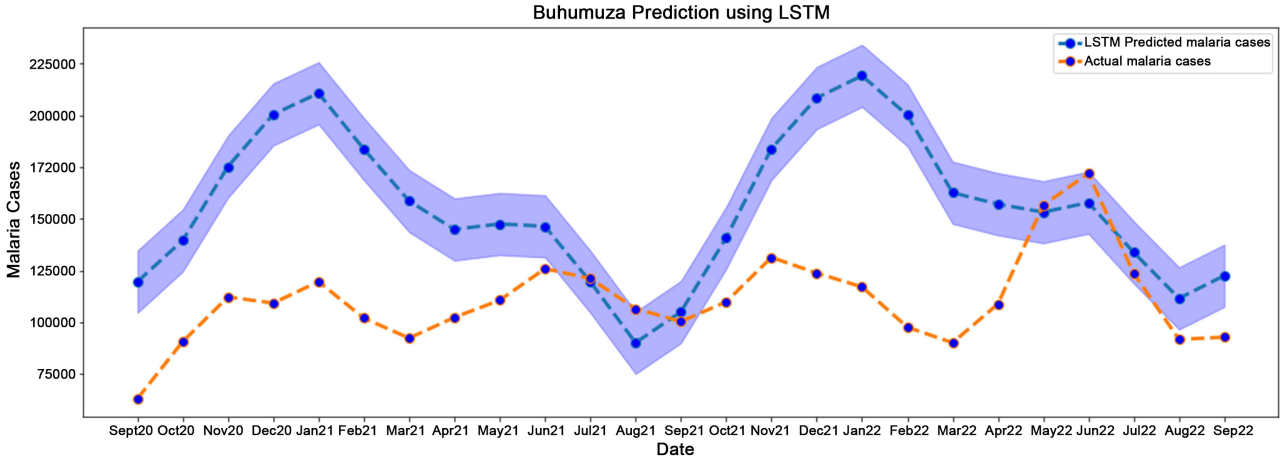


Figure 11. Multivariate LSTM Buhumuza Prediction.

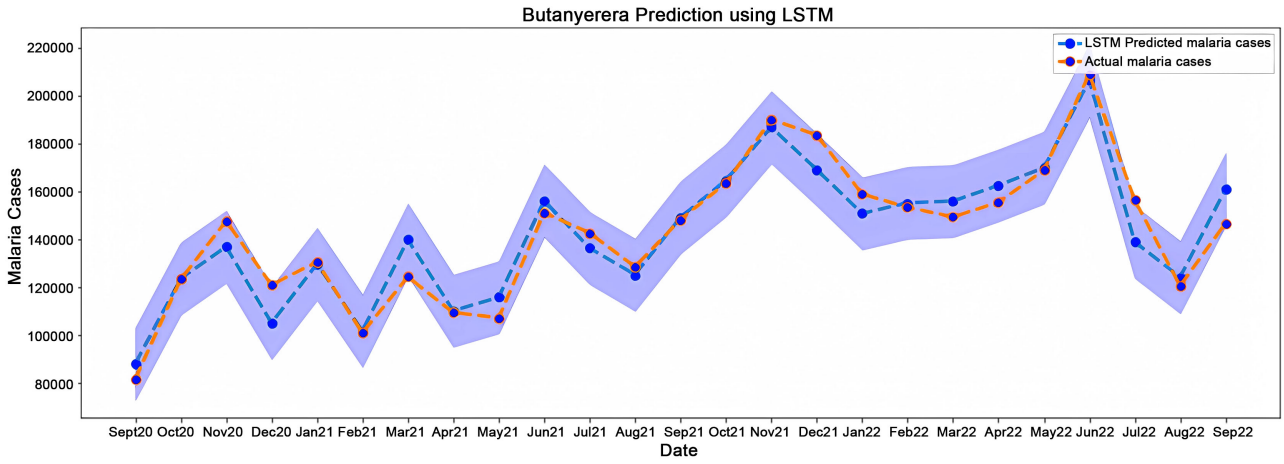


Figure 12. Univariate LSTM Butanyerera Prediction.

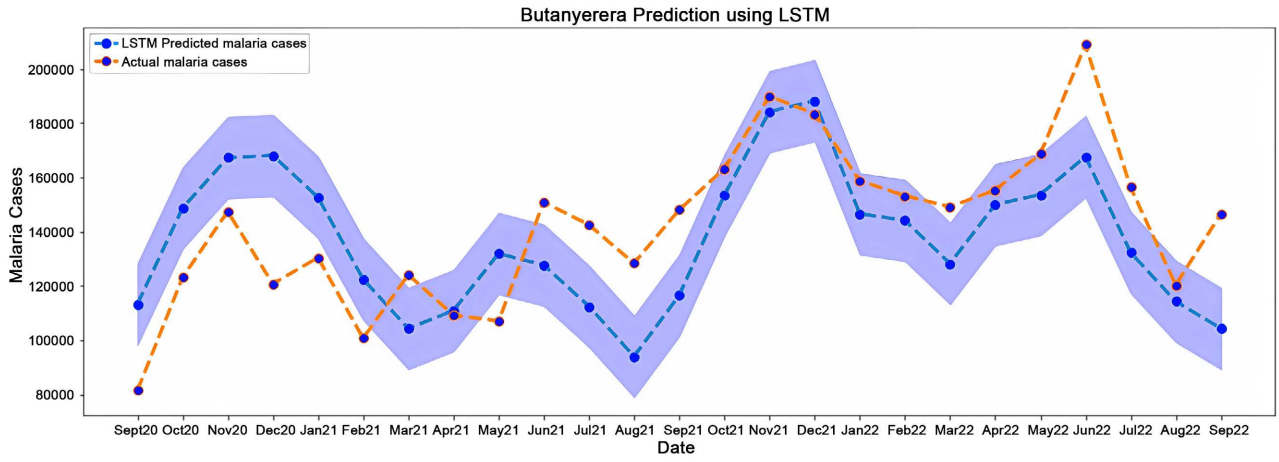


Figure 13. Multivariate LSTM Butanyerera Prediction.

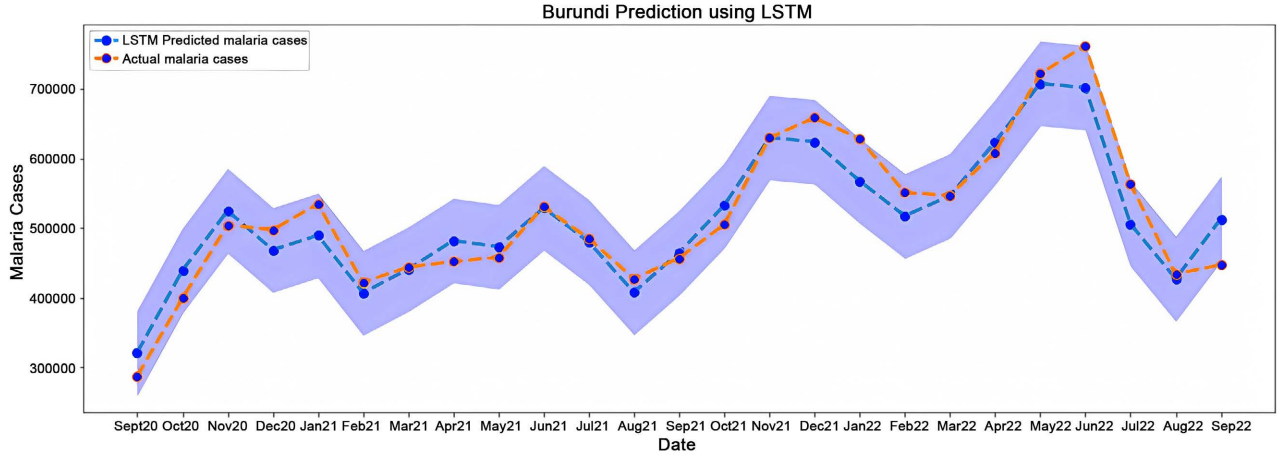


Figure 14. Univariate LSTM Burundi Prediction.

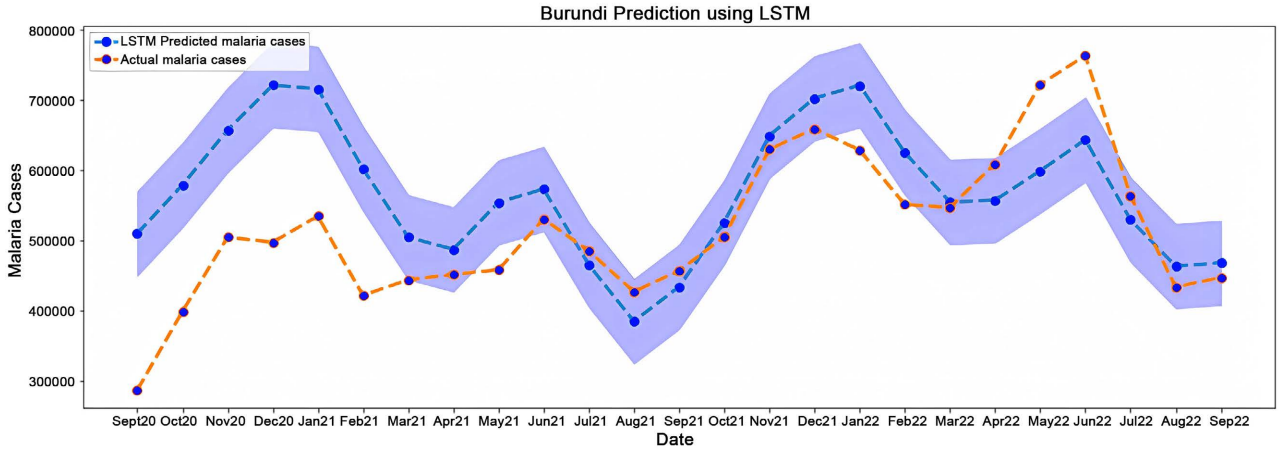


Figure 15. Multivariate LSTM Burundi Prediction.

sum for the human population and malaria cases. The curve trends were coordinated with the actual trend in all models. Nevertheless, the multivariate predicted more malaria cases than observed one while the univariate predicted less.

In comparison to the univariate model prediction, there were approximately 100,000 more malaria cases reported in the nation during that time period. However, compared to actual reported instances, the multivariate LSTM model projected a much higher number of cases, more than 2.5 million cases, than what was actually reported as seen in Table 2.

4. Discussion

In general, the multivariate LSTM model predicted more malaria cases at the country level as well as at the province level. Specifically, when examining specific provinces, the multivariate LSTM tended to overpredict cases in the northeastern provinces of Butanyerera and Buhumuza. Conversely, in the southeastern provinces of Bujumbura and Burunga, the predictions aligned with the observed trend until October 2021 but started decreasing in 2022. In Gitega province, located in the central part of the country, the predicted cases of the multivariate LSTM model were significantly lower than the observed cases. However, the univariate LSTM model demonstrated the best precision, with the predicted curve following the trend at the country and province levels. The model predictions provided a confidence interval 95%, indicating that if the process was repeated, 95% of malaria cases would fall within the range defined by the lower and upper bounds, as shown in Figures 4-15. This predictive model proved valuable in forecasting potential outbreaks. However, it is crucial to emphasize the importance of up-to-date information gathering and sharing, as the accuracy of the models relies on the most current data. Figure 14, Figure 15 illustrate that the univariate LSTM results capture the overall trends of malaria cases and provide more precise estimates at the province level. While the univariate model outperformed the multivariate model, this aligns with previous research findings [25] suggesting that the univariate model tends to excel in short term predictions, while multivariate models perform better with longer prediction horizons. This could be due to the fact that the meteorological data may not directly influence the outbreak, but over time it can impact the environmental factors that contribute to the outbreak. The consistent trends observed across the models, particularly at the country level, confirm the significant influence of climate conditions on malaria outbreaks within the country over the long term. Combining the models could improve prediction accuracy at different time steps, as noted in [26]. The confidence interval for malaria cases utilized in this study enables researchers, policymakers, and healthcare professionals to assess plausible value ranges and make informed decisions based on the estimated range. Quantifies the uncertainty associated with the estimate and provides a measure of precision of the findings. Furthermore, investigating the effects of climate change on disease outbreaks extends beyond the dynamics of malaria transmission. This study integrated meteorological factors with historical data such as human population and total monthly malaria cases. Future work could focus on daily predictions using multivariate models to explore the possible impacts of

climatic conditions, as previous studies have demonstrated the superiority of multivariate models in very short-term predictions [27]. Despite artificial neural networks often being perceived as black boxes, the results obtained highlight the importance of meteorological data in outbreak prediction, particularly at the country level where the highest number of cases were predicted. The use of deep learning models such as recurrent neural networks shows promise in predicting malaria outbreaks, which continue to pose challenges in sub-Saharan Africa. Expanding the application of artificial intelligence will facilitate collaboration among different intervention teams in the healthcare system, especially in responding to predicted increases or decreases in malaria cases.

Declarations

National data was used for this study.

A preprint has previously been published [2306.02685] [28].

Conflicts of Interest

The authors have no conflicts of interest to disclose in the publication of this paper.

References

[1] Sajana, T. and Narasingarao, M. (2017) Machine Learning Techniques for Malaria Disease Diagnosis—A Review. *Journal of Advanced Research in Dynamical and Control Systems*, **9**, 349-369.

[2] Medecins Sans Frontieres (2017) More about Malaria in Burundi.

[3] World Health Organization (WHO) (2022) World Malaria Report 2021.

[4] World Health Organization (WHO) (2017) World Malaria Report 2016.

[5] World Health Organization (WHO) (2018) World Malaria Report 2017.

[6] Institute for Health Metrics and Evaluation (2020) Health Metrics for Burundi. http://www.healthdata.org/burundi

[7] World Health Organization (WHO)/Regional Office for Africa (2017) Weekly Bulletin on Outbreaks and Other Emergencies: Week 27: 1-7 July 2017.

[8] United Nations for Children (UNICEF)/Burundi (2017) UNICEF Burundi Humanitarian Situation Report—31 March 2017.

[9] Lok, P. and Dijk, S. (2019) Malaria Outbreak in Burundi Reaches Epidemic Levels with 5.7 Million Infected This Year. *BMJ*, **366**, L5104. https://doi.org/10.1136/bmj.l5104

[10] World Vision (2017) Eight Facts about Burundi's Malaria Epidemic.

[11] USAID (2022) Burundi Malaria Operational Plan Fiscal Year 2021.

[12] Wiemken, T.L. and Kelley, R.R. (2020) Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, **41**, 21-36. https://doi.org/10.1146/annurev-publhealth-040119-094437

[13] Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., *et al.* (2006) Machine Learning in Bioinformatics. *Briefings in Bioinformatics*, **7**, 86-112. https://doi.org/10.1093/bib/bbk007

[14] Chekol, B.E. and Hagras, H. (2018) Employing Machine Learning Techniques for the Malaria Epidemic Prediction in Ethiopia. 2018 10*th Computer Science and Electronic Engineering* (*CEEC*), Colchester, 19-21 September 2018, 89-94. https://doi.org/10.1109/ceec.2018.8674210

[15] Masinde, M. (2020) Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data. *Proceedings of the* 2020 4*th International Conference on Compute and Data Analysis*, Silicon, 9-12 March 2020, 29-37. https://doi.org/10.1145/3388142.3388158

[16] Nkiruka, O., Prasad, R. and Clement, O. (2021) Prediction of Malaria Incidence Using Climate Variability and Machine Learning. *Informatics in Medicine Unlocked*, **22**, Article ID: 100508. https://doi.org/10.1016/j.imu.2020.100508

[17] Sajana, T. and Narasingarao, M. (2018) An Ensemble Framework for Classification of Malaria Disease. *ARPN Journal of Engineering and Applied Sciences*, **13**, 3299-3307.

[18] Li, D. and Huang, X. (2020) An Improved Deep Learning Model for Predicting DNA Sequence Function. *Intelligent Information Management*, **12**, 36-42. https://doi.org/10.4236/iim.2020.121003

[19] Mfisimana, L.D., Nibayisabe, E., Badu, K. and Niyukuri, D. (2022) Exploring Predictive Frameworks for Malaria in Burundi. *Infectious Disease Modelling*, **7**, 33-44. https://doi.org/10.1016/j.idm.2022.03.003

[20] Python (2018) Python Version 3.6.5.

[21] Stekhoven, D.J. (2015) Missforest: Nonparametric Missing Value Imputation Using Random Forest. Astrophysics Source Code Library.

[22] Gouvernement du Buvundi (2023) Loi organique N° 1/05 du 16 mars 2023 portant d'etermination et d'elimitation des provinces, des communes, des zones, des collines ou quartiers de la R' epublique du Burundi, Gitega.

[23] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[24] Pathan, S.M.K. and Imran, S.B. (2024) Integrated Machine Learning and Deep Learning Models for Cardiovascular Disease Risk Prediction: A Comprehensive Comparative Study. *Journal of Intelligent Learning Systems and Applications*, **16**, 12-22. https://doi.org/10.4236/jilsa.2024.161002

[25] Chayama, M. and Hirata, Y. (2016) When Univariate Model-Free Time Series Prediction Is Better than Multivariate. *Physics Letters A*, **380**, 2359-2365. https://doi.org/10.1016/j.physleta.2016.05.027

[26] Salehi, S., Kavgic, M., Bonakdari, H. and Begnoche, L. (2024) Comparative Study of Univariate and Multivariate Strategy for Short-Term Forecasting of Heat Demand Density: Exploring Single and Hybrid Deep Learning Models. *Energy and AI*, **16**, Article ID: 100343. https://doi.org/10.1016/j.egyai.2024.100343

[27] Mandal, A.K., Sen, R., Goswami, S. and Chakraborty, B. (2021) Comparative Study of Univariate and Multivariate Long Short-Term Memory for Very Short-Term Forecasting of Global Horizontal Irradiance. *Symmetry*, **13**, Article 1544. https://doi.org/10.3390/sym13081544

[28] Sakubu, D., Sinigirira, K.J.G. and Niyukuri, D. (2023) Predicting Malaria Dynamics in Burundi Using Deep Learning Models. arXiv: 2306.02685