

RESEARCH ARTICLE



OPEN ACCESS

Received: 05-05-2024

Accepted: 14-08-2024

Published: 30-12-2024

Citation: Ndao A, Dath CAB, Seck CT, Diop B (2024) Comparison of Logistic Regression Models, Random Drills and Support Vector Machines (SVM) for the Epidemiological Surveillance of Ten (10) Infectious Diseases in Senegal. Indian Journal of Science and Technology 17(48): 5083-5090. <https://doi.org/10.17485/IJST/v17i48.1512>

* **Corresponding author.**cheikhamadou.dath@ucad.edu.sn**Funding:** None**Competing Interests:** None

Copyright: © 2024 Ndao et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indjst.org/))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Comparison of Logistic Regression Models, Random Drills and Support Vector Machines (SVM) for the Epidemiological Surveillance of Ten (10) Infectious Diseases in Senegal

Abdourahmane Ndao¹, Cheikh Amadou Bamba Dath^{2*}, Cheikh Tidiane Seck¹, Boly Diop³

¹ Department of Mathematics, Alioune Diop University, Bambey, Senegal

² Atoms Lasers Laboratory, University Cheikh Anta Diop (UCAD), Dakar, Senegal

³ Prevention Department, Head of Vaccine Surveillance and Response Division, Ministry of Health, Dakar, Senegal

Abstract

Objective: The objective of this study is to analyze data from epidemiological surveillance and take into account propagation factors to track and monitor 10 infectious diseases in Senegal which are: Covid-19, Measles, Poliomyelitis (AFP), Dengue, Meningitis, Rift Valley Fever (Rift), Crimean-Congo Hemorrhagic Fever (CCHF), Chikungunya (CHIK), West Nile virus fever (WN) and Yellow Fever (YF). **Method :** The methodology is based on a comparative analysis of the performances of three supervised learning methods: logistic regression, random forests and Support Vector Machines (SVM). Data visualization methods such as Multiple Correspondence Analysis (MCA) and similarity index (Cramer's v) are also used. The data analyzed covered the period from January 2018 to November 2022 and amounted to 113,847 suspected cases. **Findings:** The results show that the methods have almost similar performances with regard to the choice of the number of selected parameters (4 variables) but differ in terms of total classification error rates, with a rate of 2.73% lower for forests random, 3.10% for logistic regression and 3.32% for Support Vector Machines. Advanced age ([40, 60] years) is identified as a direct risk factor in relation to the frequency of distribution of confirmed cases of the 10 infectious diseases. Other indirect risk factors linked to the distribution of confirmed cases are also identified and are linked to the physical environment of health districts and climatic factors. The results highlighted the ability of the random forest method to analyze surveillance data and identify complex relationships between risk factors and the distribution of infectious diseases. This study contributes to improving infectious disease surveillance and shows the importance of integrating environmental factors, socioeconomic parameters, as well as behavioral and climatic factors into data and analyses. **Novelty:** The results obtained made it possible to propose

an algorithm based on the HCPC method which combines the techniques of Principal Component Analysis (PCA) and those of Ascending Hierarchical Classification (CAH), with the application of the principle of paragons.

Keywords: Epidemiological data; Health surveillance; Logistic regression; Random Forests; Support Vector Machine

1 Introduction

Data from many domains has evolved in terms of availability, going from a few hundred pieces of data to millions. This development is particularly concerned with the field of health, among others. In this context, to process larger, more diverse, and more complex datasets, machine learning can have many implications in analytical studies that seek to identify risk factors and/or predict the resurgence of infectious diseases by taking into account many interrelated factors.

Authors have used machine learning to analyze and identify risk factors for infectious diseases from available data, in order to better understand their conditions of transmission and dissemination of infectious diseases, and therefore to better prevent them⁽¹⁾.

Some 25 baseline variables from 490 Covid-19 patients admitted to 8 hospitals in Germany were analyzed to compare the performance of 2 machine learning methods (Support Vector Machine [SVM] and Random Forest [RF]) with logistic regression⁽¹⁾. Based on the dataset obtained from Vaishnavi Clinical Laboratory, Vignesh Clinical Laboratory, Keerthi X-Ray, and Karthik Laboratory in Kumbakonam, Tamil Nadu, India, the authors compared the predictive performance of logistic regression with classification techniques such as LDA, CART, KNN, SVM, and random forest RF for predicting risk factors⁽²⁾. Vector indices and weather data are used as predictors to predict the transmission rate of dengue fever by comparing the performance of seven machine learning algorithms⁽³⁾.

Many modeling techniques have been developed for the management of infectious diseases⁽⁴⁾. Diseases are complex phenomena that depend on multiple factors, such as genetics, environment, lifestyle and exposure. To understand how these factors influence the likelihood and severity of disease, researchers most often use machine learning based only on available surveillance data.

Machine learning algorithms, including Random Forest, Logistic Regression, SVM, and k-Nearest Neighbors, were leveraged to predict the occurrence of diseases in the province of Marinduque, focusing on hand, foot and mouth disease, dengue, typhoid, influenza, chikungunya, rabies, measles, meningitis, hepatitis and acute bloody diarrhea using data from 2015 to 2019⁽⁵⁾.

The problem with using these analysis methods is that the data needed for machine learning is scarce, mainly due to privacy issues and poorly structured medical databases⁽⁶⁾. For these reasons, it is necessary to find suitable and useful sources of information and data in order to be able to build reliable prediction models from them^(7,8).

Data relating to certain risk factors for spread are not always available in the databases of infectious diseases under surveillance in Senegal. The available monitoring data is incomplete and often subject to errors. However, there are limits to take into account in the national health surveillance system. Hence this is what justifies in this context that we find the interest of supplementing data from the integrated surveillance of 10 infectious diseases which are: Covid-19, Measles, Poliomyelitis (AFP), Dengue, Meningitis, Rift Valley Fever (Rift), Crimean-Congo Hemorrhagic Fever (CCHF), Chikungunya (CHIK), West Nile Virus Fever (WN) and Yellow Fever (YF), in order to have a reliable database for all statistical studies and usable via emerging technologies.

In this work, we compare the performance of a supervised statistical learning method, namely logistic regression, with two machine learning methods: random forests and Support Vector Machines (SVM), to analyze the surveillance data for these 10 infectious diseases with a view to supplementing them with the identification of risk factors for spread. After the summary and the introduction, the rest of the work is organized as follows: Section 2 describes the methodology as well as the tools used. Section 3 presents and discusses the results. Section 4 concludes the paper.

2 Methodology

2.1 Data

We worked with the global database with 60% of confirmed positive cases compared to 40% of confirmed negative cases. The data are extracted from Base_RR_2018_2022 (Measles database, DP-MSAS), LCR HEAR (Meningitis database, Albert Royer Children's Hospital - DAKAR), PFA_2018-2022_S-48 (PFA database, DP-MSASS), Base-Covid-19-Senegal_2020-2022 (Covid-19 database, COUS-MSASS) and BDD_4S_MSAS_19122022 (YF, Rift, Dengue, CHIK, CCHF, WN, Institute database Pastor of Dakar).

The statistics recorded in the available databases concern: the result of suspected cases; the health district of residence (DS_Resids); the medical region of residence (RM_Resids); age (grouping); genre; the epidemiological week (Week); the start date of the signs (Date.Sign) and the disease in question (Caract_Mald). The database is split in two randomly, with on one side 90% of the observations used to estimate the model parameters, and on the other hand 10% for a model validation step. The training sample is used to build the methods and select their parameters, and thus the validation sample to estimate the performance of the methods.

The response variable $Y = \text{Result}$ takes two modalities: "Positive" or "Negative" and the explanatory variables are the 6 other qualitative variables. Age (grouping) is considered here as categorical because it is discretized into groups. The modalities of the DS_Resids and Week variables have been grouped together because some are poorly represented in the training sample.

Selection criteria for the ten infectious diseases:

Senegal has set a measles elimination threshold of less than 1 case per 1 million inhabitants. However, a dramatic increase in measles cases has been observed over the last 5 years with a national incidence of 0.8 in 2018 rising to 24.3 in 2023⁽⁹⁾. In 2021, an epidemic of circulating vaccine-derived poliovirus type 2 (cVDPV2) reappeared in Senegal while the last case of wild poliovirus dates back to 2010. A total of 18 human cases and 15 environmental isolates of cVDPV2 were recorded⁽¹⁰⁾. From March 2, 2020 to December 31, 2022, Senegal reported 14 984 Covid-19 patients currently hospitalized, including 138 who died⁽¹¹⁾. For measles, out of the 1115 confirmed cases during the period from January 2018 to November 2022, 30 have died (Base_RR_2018_2022: Measles database, DP-MSAS).

An analysis based on confirmed cases shows that in recent years, Senegal has experienced a high incidence of epidemics with infectious diseases, some of which were in the process of being eliminated or eradicated (AFP, Measles) and others emerging due to climate change (Dengue, Rift Valley Fever, Crimean Congo Hemorrhagic Fever, Covid-19)⁽¹²⁾. These diseases are among the priority diseases monitored by the health system as part of Integrated Disease Surveillance and Response (IDSR). Indeed, given their reappearance, these diseases have become priorities for the national health monitoring system. These observed phenomena also justify the selection of these 10 diseases subject to epidemiological surveillance in Senegal through the SIMR strategy. It is in this context that we were able to analyze the available data by supplementing them with the risk factors associated with the spread.

2.2 Method

The method is based on a comparative analysis of the performance of three supervised learning methods: logistic regression, random forests and Support Vector Machines (SVM) and data visualization tools such as Multiple Correspondence Analysis (ACM) and the similarity index (Cramer's v).

Logistic regression is a powerful tool in public health that not only predicts outcomes but also identifies risk factors influencing the development of infectious diseases. In addition, the random forest algorithm excels in analyzing complex data. Its ability to handle large datasets and select relevant variables makes it a leading choice for epidemiological surveillance. With the explosion of available data and the need for accurate and fast analyses, support vector machines (SVMs) are also being implemented. SVMs represent a robust and flexible approach in the field of public health, allowing to efficiently process complex data and improve the decision-making process.

Hyperparameter selection and model validation: we use a backward algorithm for model parameter selection. For SVMs, the selection is performed using the SVM-RFE algorithm with a Gaussian kernel. In the context of random forests, we use the RFE (Recursive Feature Elimination) procedure. To guide this selection, we use the permutation importance measure of the variables. As for the logistic regression, the choice of variables is made using the AIC (Akaike's Information Criterion) criterion. We evaluate the performance of the models using the Out Of Bag (OOB) technique, which estimates the prediction error thresholds using the validation data (Table 3). The Out Of Bag approach consists of using the observations that are not included in the bootstrap samples as a validation sample to evaluate the performance of the models. This method avoids the need for cross-validation while providing an accurate estimate of the performance of the models in terms of prediction.

The fitted random forest with $mtry = 2$ and $ntree = 500$ shows the lowest prediction error threshold (2.73%). We prefer this Out-of-bag method because it is better suited to random forests.

Machine learning-based analysis methods have proven effective in predicting the spread of diseases and improving decision-making for future interventions⁽¹³⁾. Different forecasting methods are widely used to solve forecasting problems. The comparative performance analysis proposed in this study aims to select the best performing algorithm to analyze and identify significant risk factors associated with the spread of infectious diseases, by exploiting high-dimensional epidemiological surveillance data. To verify that these epidemiological data are reliable, accurate, and ready to be used in analyses and models for public health decision-making, we proceed to the following steps:

- Preprocess the available data using a data cleaning, variable transformation and data structuring process under R version 4.3.3 (2024-02-29 ucrt).
- Explore the relationships between qualitative or discretized explanatory variables using Multiple Correspondence Analysis (MCA). This technique also included the response variable, thus making it possible to verify the absence of correlation between the variables on the factorial level. The analysis revealed a strong correlation between the spatial variables DS_Resids and RM_Resids. This result guided our selection of the DS_Resids variable for the steps of comparing the performances of the three methods.
- Evaluate the intensity of the link between the response variable and the qualitative or categorical variables using the Cramer V similarity index (normalized χ^2), the results of which are presented in Table 1. The values of this index reveal that the variables Caract_Mald (Cramer $v = 0.94$) and DS_Resids (Cramer $v = 0.73$) are very strongly associated with confirmed cases of the 10 infectious diseases. The variables Week (Cramer $v = 0.43$) and grpage (Cramer $v = 0.43$) also show a significant link with the target variable Result. The variable sex of the suspected case has low significance (Cramer's $v = 0.02$) in explaining confirmed cases of these 10 infectious diseases.

Table 1. Similarity matrix (Cramer's v)

	DS_Resids	grpage	Caract_Mald	Gender	Week	Result
DS_Resids	1.00	0.16	0.27	0.07	0.20	0.73
Grpage	0.16	1.00	0.30	0.09	0.10	0.43
Caract_Mald	0.27	0.30	1.00	0.09	0.19	0.94
Gender	0.07	0.09	0.09	1.00	0.05	0.02
Week	0.20	0.10	0.19	0.05	1.00	0.43
Result	0.73	0.43	0.94	0.02	0.43	1.00

The 4 variables (Table 2) "DS_Resids, grpage, Week, Caract_Mald" are therefore kept to compare the performances of these three supervised learning algorithms.

Table 2. Presentation of selected variables

Variable	Modality	Status	Description
Result	Figure 1	Target	Positive cases following laboratory diagnostic tests
DS_Resids	Figure 1	Active	Health districts of residence of suspected case (epidemiological unit)
grpage	Figure 1	Active	Age range of suspected cases of the 10 infectious diseases
Week	Figure 1	Active	Epidemiological week from January 2018 to November 2022
Caract_Mald	Figure 1	Active	Characteristics of the disease in question detected in the patient

The selected variables (Table 2) are all nominal, and we present, in Figure 1, the modalities of the analysis variables.

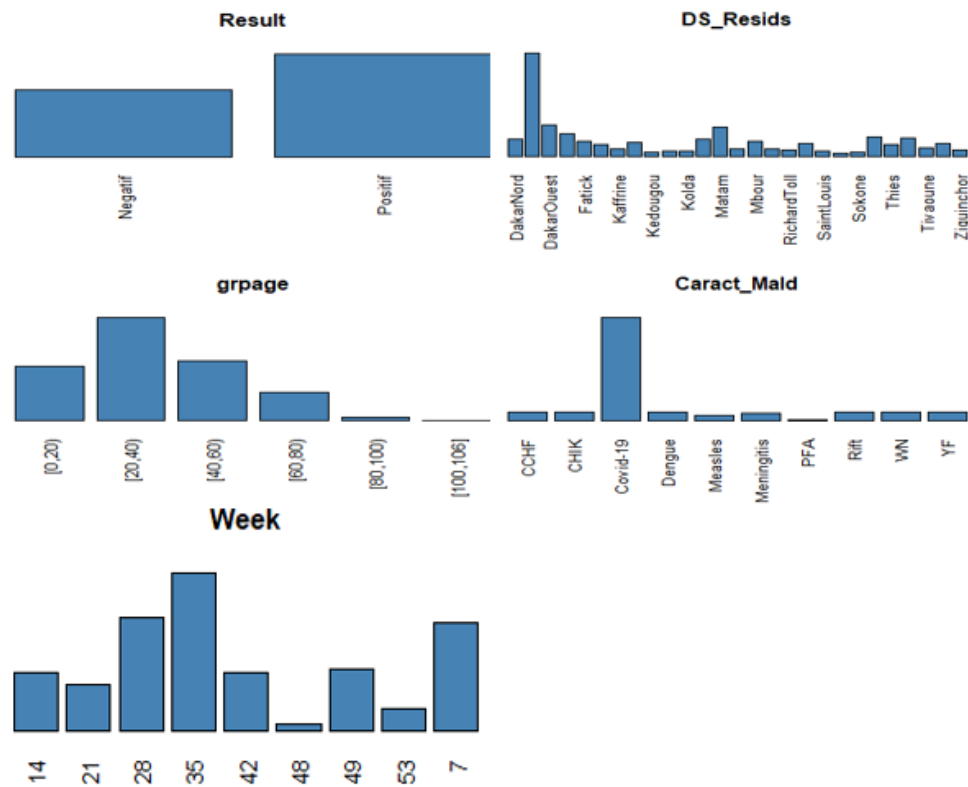


Fig 1. Frequency display for selected variable categories

3 Results and Discussion

The results show at the $\alpha = 5\%$ threshold that the tests are significant for the 4 variables retained (Table 3). The three methods adjusted to the data indicate that the parameters linked to health districts (DS_Resids), to the intrinsic characteristics of the 10 diseases (Caract_Mald), to climate (Week) and to advanced age: grouping ([40, 60] years) are risk factors that may influence the frequency of distribution of confirmed cases of these 10 infectious diseases within the 79 health districts.

Comparative analysis of classification models reveals that random forests stand out for their ability to select a more parsimonious set of explanatory variables. With only four variables selected, random forests have the lowest total misclassification rate of 2.73%. This exceptional performance is also reflected in the area under the ROC curve, reaching 97.3%, which is the best result among the three models evaluated (Table 3).

Table 3. Number of selected variables, error rate and area under the ROC curve of the three models

Méthodes	Selected variables	Error rate	AUC
Logistic regression (LR/AIC)	4	3.10%	96.16%
Random forests (RF/RFE)	4	2.73%	97.30%
SVM – REF	4	3.32%	96.90%

Error rates and Receiver Operating Characteristic (ROC) curves are key elements in the evaluation and improvement of epidemiological surveillance systems. They ensure that public health decisions are based on reliable data, thus contributing to effective management of health crises. Error rates in epidemiological surveillance refer to the proportion of false positives and false negatives in the available data. ROC curves plot the true positive rate (sensitivity) against the false positive rate, at different decision thresholds. They help determine the optimal threshold for classifying an event as positive or negative. This is important in epidemiological surveillance, where poorly defined thresholds can lead to classification errors. The error rates estimated by the models are relatively very low, while the area under the ROC curves reaches very high values (Table 3), thus highlighting the effectiveness of the national epidemiological surveillance system. In the context of epidemiological surveillance, it is essential

to provide more precision on the performance measures of predictive models in order to evaluate their effectiveness in disease detection and monitoring. For this purpose, we present in Table 4 the metrics such as accuracy (the prediction probability threshold), precision, recall and F1-score.

Table 4. Performance Metrics of the Three Models in Epidemiological Surveillance

Methods	Accuracy	Precision	Recall (Sensibility)	F1-Score
Logistic regression (LR/AIC)	96.90%	96.83%	97.84%	97.33%
Random forests (RF/RFE)	97.27%	97.25%	95.12%	96.17%
SVM	96.68%	96.41%	95.58%	95.99%

A comparative analysis based on performance metrics (Table 4), confirms that random forests outperform both support vector machines (SVM) and logistic regression (LR). Our results align with those of⁽¹⁴⁾ who performed a similar analysis comparing these three classification algorithms to predict coronary heart disease (CHD). Their study, based on data from 350 patients collected in a hospital in Armenia, observed a high accuracy of 95% and suggested a ranking of the three algorithms studied according to their performance as follows: Random Forest (RF) comes first in terms of performance, followed by the Support Vector Machine (SVM) method and then the Logistic Regression (LR) methods. These results reinforce the idea that random forests are well suited as a tool for epidemiological surveillance, offering both high accuracy and the ability to efficiently handle complex data.

Other studies have compared different machine learning models with logistic regression to analyze epidemiological surveillance data. For example, a study⁽¹⁵⁾ on predicting the risk of lymph node metastasis in gastric cancer patients found that among seven algorithms tested, except SMV, the other models had higher accuracy and reliability than logistic regression. Another study⁽¹⁶⁾ aimed at identifying the severity factors of Covid-19 at admission, comparing the predictive performance of random forests (RF) and support vector machines (SVM) with logistic regression (LR).

The results showed that SVM outperformed LR in terms of accuracy (88.5% vs. 85.2%), followed by RF (84.5%). Unlike our study, we observed that random forests (RF) were able to effectively analyze data from ten infectious diseases, displaying higher accuracy (97.25%) compared to logistic regression (96.83%) and SVM (96.41%). Our results highlight the ability of random forests (RF) to analyze data from integrated infectious disease surveillance. However, to optimize the use of random forests as an analysis tool in the national health surveillance system, it is essential to complete the available data in order to establish a more comprehensive database, that integrates the risk factors associated with the targeted infectious diseases. The results of our study showed that the random forest algorithm achieved the highest level of predictive accuracy at 97.27%. This performance confirms the observations of⁽¹⁷⁾, who highlighted that the evaluation of machine learning models often focuses on predictive accuracy, and highlights that the study of diseases requires a high rate of correct classification of patients.

In our study, the application of random forests made it possible to distinguish two types of explanatory risk factors for confirmed cases of the 10 infectious diseases: on the one hand, indirect risk factors related to health districts (variable: DS_Resids), intrinsic disease characteristics (variable: Caract_Mald) and climate (variable: Week); on the other hand, direct risk factors associated with age ([40, 60] years). In a previous study,⁽¹⁸⁾ used data from 3 662 Iranian Army ground forces to compare three supervised learning methods: logistic regression, decision tree, and random forest, to study the association of type 2 diabetes mellitus (T2DM) with risk factors and classify subjects in a military population. The results confirmed the superiority of random forest in identifying complex relationships between risk factors, with significantly improved sensitivity and specificity. The ROC curve showed higher performance for random forest (97.1%) compared to logistic regression (73.8%) and decision tree (77.1%). These results also highlighted the importance of variables such as age, body mass index, total cholesterol, low-density lipoprotein cholesterol, and triglycerides in predicting T2DM risk. Some of these variables, including age, were also identified as significant risk factors in our study for confirmed cases of the 10 infectious diseases studied.

The study conducted by⁽¹⁹⁾ compared the performance of three supervised machine learning techniques: linear discriminant analysis (LDA), random forest (RF), and support vector machine (SVM) on Covid-19 datasets based on demographic, clinical, and epidemiological characteristics. Their results showed that random forest was the best performing algorithm, with a prediction accuracy of 100%, compared to 95.2% for LDA and 90.9% for SVM. In addition, the chi-square test revealed that seven features, except gender, were significantly correlated with the outcome of Covid-19 patients. Studies^(11,18) have also highlighted age as a determining factor associated with poor prognosis (death) in Covid-19 patients. These results^(11,18,19) are consistent with the findings of our own study. Many other studies have also shown, through statistical or machine learning approaches, that age and comorbidities are significant risk factors for the spread of infectious diseases; see for example^(20–23).

The results of the study⁽⁵⁾ clearly revealed that the Random Forest algorithm was the most accurate among those evaluated, which is of great importance for infectious disease surveillance and control and efficient resource allocation (Table 5).

Table 5. Predictive Performance of Machine Learning Algorithms⁽⁵⁾

Algorithm	Accuracy	precision	recall	F1-score
Random Forests	70.12%	68.06%	70.12%	68.65%
Logistic regression	54.42%	29.62%	54.42%	38.36%
SVM	51.26%	31.54%	53.11%	43.45%
K-nearest neighbors	63.41%	60.54%	63.41%	60.61%

The results of our study are consistent with those available in some literature, confirming that random forest (RF) displays the best performance for the analysis of data related to the 10 infectious diseases studied⁽⁵⁾. Our analysis reveals an accuracy of 97.25% for random forest, surpassing the performance of logistic regression (LR) at 96.83% and support vector machines (SVM) at 96.41%. These results mark a significant improvement in the field. Some common infectious diseases identified as subject to the analysis of risk factors of spread in our study and in some literature include dengue, chikungunya, measles and meningitis⁽⁵⁾. This comparative analysis of the performance of random forest and SVM with logistic regression from the surveillance data of the 10 infectious diseases has limitations. The large sample size of 113,847 suspected cases, not including all risk factors for spread, could impact the reliability of the epidemiological analysis. This highlights that the available data is incomplete and that the identification of direct and indirect risk factors is necessary to complete the available data and, finally, to establish a reliable database for all statistical studies. In this context, additional multifactorial studies using this complete database will be essential to confirm our results. These studies will aim to strengthen the validity of the conclusions drawn from RF, SVM and LR models in predicting the resurgence of infectious diseases, by taking into account the risk factors significantly associated with the spread.

4 Conclusion

This study presents a novel application of machine learning methods to epidemiological surveillance data in Senegal, comparing logistic regression, random forests, and SVMs for ten infectious diseases. Our results indicate that random forest was the best performing model for analyzing, preventing, monitoring, and predicting infectious disease outbreaks.

The originality of this study lies in the integration of machine learning methods, particularly random forest, within the national health surveillance system. To strengthen this approach, it is essential to improve data collection by integrating environmental, behavioral, climatic, socioeconomic factors, and disease-specific characteristics. This approach aims to optimize models and improve forecast accuracy.

By integrating the random forest method, the national health surveillance system could benefit from more effective response strategies and an improved ability to anticipate infectious disease outbreaks. These advances are essential to guide the actions of health professionals and political decision-makers to optimize responses to epidemic crises in time.

References

- 1) Sievering AW, Wohlmuth P, Gefßler N, Gunawardene MA, Herrlinger K, Bein B, et al. Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission. *BMC Medical Informatics and Decision Making*. 2022;22(1). Available from: <https://dx.doi.org/10.1186/s12911-022-02057-4>.
- 2) Kannan K, Menaga A. Risk Factor Prediction by Naive Bayes Classifier, Logistic Regression Models, Various Classification and Regression Machine Learning Techniques. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences*. 2022;92(1):63–79. Available from: <https://dx.doi.org/10.1007/s40011-021-01278-3>.
- 3) Ong SQ, Isawasan P, Ngesom AMM, Shahar H, Lasim AM, Nair G. Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data. *Scientific Reports*. 2023;13(1). Available from: <https://dx.doi.org/10.1038/s41598-023-46342-2>.
- 4) Khaddaj S, Chrief H. Prevention and Control of Emerging Infectious Diseases in Human Populations. In: 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). IEEE. 2020;p. 336–339. Available from: <https://ieeexplore.ieee.org/document/9277827>.
- 5) Go VDM. Communicable disease surveillance through predictive analysis: A comparative analysis of prediction models. *HO CHI MINH CITY OPEN UNIVERSITY JOURNAL OF SCIENCE - ENGINEERING AND TECHNOLOGY*. 2023;13(2):45–54. Available from: <https://dx.doi.org/10.46223/hcmcoujs.tech.en.13.2.2944.2023>.
- 6) Aftab K, Fatima HS, Aziz N, Baig E, Khurram M, Mubarak F, et al. Machine Learning and Sampling Techniques to Enhance Radiological Diagnosis of Cerebral Tuberculosis. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET). IEEE. 2021;p. 1–5. Available from: <https://ieeexplore.ieee.org/document/9659603>.
- 7) Kim J, Ahn I. Infectious disease outbreak prediction using media articles with machine learning models. *Scientific Reports*. 2021;11(1). Available from: <https://dx.doi.org/10.1038/s41598-021-83926-2>.
- 8) Roster K, Connaughton C, Rodrigues FA. Forecasting new diseases in low-data settings using transfer learning. *Chaos, Solitons & Fractals*. 2022;161. Available from: <https://dx.doi.org/10.1016/j.chaos.2022.112306>.

- 9) Direction Générale de la Sante Publique / Direction de la Prevention. 2024. Available from: <https://www.sante.gouv.sn/les-directions/la-direction-g%C3%A9n%C3%A9rale-de-la-sant%C3%A9-publique>.
- 10) Mbaye AH. Evaluation des performances de la campagne de rattrapage au VPI, Mémoire de Suivi et Evaluation. 2021.
- 11) Seck CT, Faye I, Diop A, Niang MA, Sylla SN, Ndao A, et al. Prognosis Method on the Outcome of Covid-19 Patients in Senegal. *African -British Journal*. 2023;6(3):93–103. Available from: <https://abjournals.org/ajmss/papers/volume-6/issue-3/prognosis-method-on-the-outcome-of-covid-19-patients-in-senegal/>.
- 12) Msas/Dp. Guide technique nationale pour la surveillance intégrée des maladies et riposte. 2019. Available from: <https://iris.who.int/bitstream/handle/10665/331254/WHO-AF-WHE-06-2019-fre.pdf?sequence=1&isAllowed=y>.
- 13) Mohan M, Vansh Agrawal, Malik M, Atul. Monitoring Novel Corona Virus using Machine Learning. In: 2022 International Conference on Inventive Computation Technologies (ICICT). IEEE. 2022;p. 824–830. Available from: <https://ieeexplore.ieee.org/document/9850795>.
- 14) Ghorbani H, Krasnikova A, Ghorbani P, Ghorbani S, Hovhannisyan HS, Minasyan A. Improving the Estimation of Coronary Artery Disease by Classification Machine Learning Algorithm. In: IEEE 6th International Conference and Workshop Obuda on Electrical and Power Engineering (CANDO-EPE). IEEE. 2023;p. 159–166. Available from: <https://ieeexplore.ieee.org/document/10418014>.
- 15) Lu T, Fang Y, Liu H, Chen C, Li T, Lu M, et al. Comparison of Machine Learning and Logic Regression Algorithms for Predicting Lymph Node Metastasis in Patients with Gastric Cancer: A two-Center Study. *Technology in Cancer Research & Treatment*. 2024;23. Available from: <https://dx.doi.org/10.1177/15330338231222331>.
- 16) Xiong Y, Ma Y, Ruan L, Li D, Lu C, Huang L, et al. Comparing different machine learning techniques for predicting COVID-19 severity. *Infectious Diseases of Poverty*. 2022;11(1). Available from: <https://dx.doi.org/10.1186/s40249-022-00946-4>.
- 17) Sahebbonar M, Dehaki MG, Kazemi-Galougahi MH, Soleiman-Meigooni S. A Comparison of Three Research Methods: Logistic Regression, Decision Tree, and Random Forest to Reveal Association of Type 2 Diabetes with Risk Factors and Classify Subjects in a Military Population. *Journal of Archives in Military Medicine*. 2022;10(2). Available from: <https://dx.doi.org/10.5812/jamm-118525>.
- 18) Abdulhameed AO, Mannir A, Usman M, Auwalu I, Lawan IA, S AA, et al. A Classification Approach for Predicting COVID-19 Patient's Survival Outcome with Machine Learning Techniques. *Science Forum (Journal of Pure and Applied Sciences)*. 2020;p. 1–10. Available from: <https://www.ejmanager.com/mnsteps/7/7-1608137167.pdf?t=1735625205>.
- 19) Bonanad C, García-Blas S, Tarazona-Santabalbina F, Sanchis J, Bertomeu-González V, Fácila L, et al. The Effect of Age on Mortality in Patients With COVID-19: A Meta-Analysis With 611,583 Subjects. *Journal of the American Medical Directors Association*. 2020;21(7):915–918. Available from: <https://dx.doi.org/10.1016/j.jamda.2020.05.045>.
- 20) Asfahan S, Deokar K, Dutt N, Niwas R, Jain P, Agarwal M. Extrapolation of mortality in COVID-19: Exploring the role of age, sex, co-morbidities and health-care related occupation. *Monaldi Archives for Chest Disease*. 2020;90(2):313–330. Available from: <https://dx.doi.org/10.4081/monaldi.2020.1325>.
- 21) Tian W, Jiang W, Yao J, Nicholson CJ, Li RH, Sigurslid HH, et al. Predictors of mortality in hospitalized COVID-19 patients: A systematic review and meta-analysis. *Journal of Medical Virology*. 2020;92(10):1875–1883. Available from: <https://dx.doi.org/10.1002/jmv.26050>.
- 22) Yu C, Lei Q, Li W, Wang X, Liu W, Fan X, et al. Clinical Characteristics, Associated Factors, and Predicting COVID-19 Mortality Risk: A Retrospective Study in Wuhan, China. *American Journal of Preventive Medicine*. 2020;59(2):168–175. Available from: <https://dx.doi.org/10.1016/j.amepre.2020.05.002>.
- 23) Shi Q, Zhang X, Jiang F, Zhang X, Hu N, Bimu C, et al. Clinical Characteristics and Risk Factors for Mortality of COVID-19 Patients With Diabetes in Wuhan, China: A Two-Center, Retrospective Study. *Diabetes Care*. 2020;43(7):1382–1391. Available from: <https://dx.doi.org/10.2337/dc20-0598>.