MATHEMATICAL MODELING AND COMPUTING, Vol. 13, No. 1, pp. 52–59 (2026)

Resurgence Prediction of Ten Infectious Diseases under Surveillance in Senegal

Ndao A.[1], Seck C. T.[1,*], Diop B.[2]

[1] *Université Alioune Diop, BP 30 Bambey, Sénégal*
[2] *Direction de la Prévention, Ministère de la Santé, Dakar, Sénégal*
[*] *Corresponding author: cheikhtidiane.seck@uadb.edu.sn*

(Received 16 August 2025; Revised 21 January 2026; Accepted 28 January 2026)

In this paper, there are proposed two multi-class predictive models for estimating the resurgence probability of ten infectious diseases under epidemic surveillance in Senegal. The first model is a Multiple Binary Random Forest (MBRF), which utilizes the ranger function with Gini criterion and allows to separately predict each of the ten diseases while taking account of their interdependencies. The second model is a Multi-Output Decision Tree (MODT), which introduces an inertia criterion (calculated with Chi-square distance) as the node impurity measure and allows to simultaneously predict all of ten diseases. Data come from the global disease surveillance database of the Ministry of Health, and contain information, on 68 698 instances, related to disease's, district's as well as patient's characteristics. The results showed that, during the study period (January 2018 to November 2022), these ten pathologies recorded an average resurgence probability of 12.2%, except for Poliomyelitis, which had a lower score estimated at 2.4%, and Covid-19 which showed a fairly high resurgence rate hovering 60%. Compared to standard algorithms such as: multi-class random forests (MCRF) and multinomial logistic regression (MLR), our two models provided better performance. For example, for F1-score, we have: MBRF (0.9999), MODT (0.8572), MCRF (0.8451), MLR (0.8211).

Keywords: *predictive models; multi-class models; resurgence probability; infectious diseases.*

2020 MSC: 68-04, 68Q32, 68T05 **DOI:** 10.23939/mmc2026.01.052

1. Introduction

In recent decades we have witnessed major changes in our global environment: human behaviors and activities, climate change, environmental modifications, pathogens' evolution, etc. This poses a real public health concern, as many infectious diseases have emerged or reappeared due to the disruption of ecosystems. For instance, the spread of infectious diseases, such as those transmitted by tiger mosquito or ticks, is increasing worldwide. One can also note a resurgence of zoonoses in many regions of the globe.

To address this situation, the WHO (World Health Organization) has initiated a strategic plan to combat these emerging or re-emerging infections by reinforcing surveillance, alert and response, applied research, prevention and control, and strengthening public health structures particularly in developing countries. In Senegal, health authorities have adopted the Integrated Disease Strategy and Response (IDSR), as have done many other country members of the WHO African Region. This strategy aims at improving disease surveillance and strengthening response capacity to face epidemics and other health emergencies. In addition to the IDSR program, there is a network named *4S* (Syndromic Sentinel Surveillance Network in Senegal), which has been set up in 2012 thanks to the collaboration between the Ministry of Health and the Pasteur Institute of Dakar.

https://orcid.org/0009-0000-3649-6325 (Ndao A.), https://orcid.org/0000-0002-2920-5371 (Seck C. T.), https://orcid.org/0000-0001-5950-2628 (Diop B.)

A large number of diseases are included in the IDSR surveillance program, but in this work, we focus on ten of them, whose monitoring is particularly crucial for health authorities. Indeed, these latter are potentially epidemic infectious diseases and belong to the priority diseases for the national health system. Some of them are emerging and others, which were considered as eradicated a few decades ago, are reappeared. These ten pathologies are: Covid-19, Measles, Poliomyelitis (PFA), Dengue Fever, Meningitis, Rift Valley Fever (Rift), Crimean Congo Hemorrhagic Fever (CCHF), Chikungunya (CHIK), West Nile Fever (WN) and Yellow Fever (YF).

Predictive models allow the forecast of epidemic trends before they occur, facilitating thus anticipation of health systems for early and effective response. To build high-performance models, many researchers rely on a machine learning approach. For example, [1] developed an LSTM to predict dengue cases in Brazil from climatic and spatial variables and identified the most important climatic predictors using SHAP method. While [2] provided a systematic review of several machine learning algorithms, and showed the possibility of combining them in order to obtain accurate forecasts of the incidence and trends of many infectious diseases.

Our goal in this paper is to develop two machine learning models for the resurgence prediction of the ten above mentioned infectious diseases. To this end, we will use a multi-label classification approach, that presents challenges in terms of accuracy and efficiency as the number of classes to be predicted is large; see [3]. When there are several tasks to learn, [4] suggest learning them simultaneously rather than separately if the tasks are correlated; improving thus predictive performance. Also, [5] proposes a multi-output random forest model, that generalizes the tree induction algorithm of [6] and allows learning simultaneously multiple classification and regression tasks. This model is based on an impurity measure defined as a combination of Shannon and differential entropies.

Whenever there are multiple quantitative output variables which are covariant, multivariate regression trees studied in [7] generally provide good predictions. [8] applied this strategy to produce multivariate regression trees and solve classification problems for geographic and ecological data. Also, [9] utilized multivariate regression trees to build a random forest model with multiple responses. While [10] applied classification trees to analyze multiple binary responses. The common feature for all these models is that they employ an impurity measure based on covariance-weighted entropy or least square distance, which requires to determine the covariance structure of the output variables and then to ultimately work with quantitative data.

In this paper, we propose two predictive multi-output models that can deal directly with qualitative data and ignore the covariance structure required in the calculation of the impurity measure in the above models. The first model is a Multiple Binary Random Forest (MBRF) algorithm, which utilizes the ranger function with Gini criterion and allows to separately predict each class while taking account of possible interdependencies. The second model is a Multi-Output Decision Tree (MODT) algorithm, which introduces an inertia criterion, inspired from Multiple Correspondence Analysis technique, as the node impurity measure and allows to jointly predict the occurrence of all ten diseases. These two models enable us to estimate resurgence probabilities for each of the ten targeted diseases, by taking the proportion of positives predictions of each disease in the test sample.

The paper is structured as follows: Section 2 describes the methodology, Section 3 presents and discusses the results, while Section 4 concludes the work.

2. Methodology

2.1. Data

The data were extracted from the Senegalese global disease surveillance database and concern confirmed cases (patients) of ten infectious diseases, recorded during the period from January 2018 to November 2022. The studied database contains 68,698 instances or observations. Each instance is associated with a patient from a given district and provides: the patient's age and sex, the diagnosed disease and its characteristics, as well as characteristics of the health district such as: environmental and climatic conditions, human behavior, and socioeconomic living conditions. The study variables are presented in Table 1.

Table 1. Definition of analysis variables.

Variables	Description	Nature	Type	Modalities
Disease	Type of disease	Qualitative	Nominal	Covid-19, Measles, PFA, Dengue, Meningitis, Rift, CCHF, CHIK, WN, YF
ModTrans	Mode of transmission	Qualitative	Binary	Direct, Indirect
Vaccine	Existence of a vaccine	Qualitative	Binary	Yes, No
VitesProp	Speed of propagation	Qualitative	Ordinal	Fast, Moderate, Slow
FactEnv	Environmental factors	Qualitative	Nominal	Lack of hygiene& sanitation, Pollution, Presence of enclosures/parks
FactCompHum	Human behavior factors	Qualitative	Nominal	Promiscuity, High population density, Social interrelations
FactSocioEco	Socio-economic factors	Qualitative	Nominal	Poverty, Existence of public meeting places, Lack of access to healthcare
FactClimat	Climatic factors	Qualitative	Nominal	Temperature, Wind and Dust, Rain, Humidity
Rcrudes	Recrudescence period	Qualitative	Nominal	Winter season, Dry season
Grpage	Age of patients	Qualitative	categorical	[0, 20[– [20, 40[– [40, 60[– [60, 80[– [80, 106]
Gender	Patient's gender	Qualitative	Nominal	Man, Woman

2.2. Models

The aim of this paper is to simultaneously predict the resurgence of ten infectious diseases based on spread risk factors or variables defined in Table 1. We can consider this problem as a multi-label classification problem, where the output variable has more than two labels or classes. To solve such a problem, we can reduce it to a multi-output problem, i.e. where the output is a vector rather than a single element. Thus, depending on whether the output components are correlated or not, we may use either algorithms that make separate predictions with individual outputs or algorithms that make simultaneous predictions for all components of the targeted variable.

A classic and simple approach to multi-label prediction problems is the "Binary Relevance" one. But, its drawback is that it does not take into account the dependencies between labels. To overcome this difficulty, some approaches such as: Classifier Chains; see, e.g., [11] and Classifier Treillis; see, e.g., [12] have been proposed. But, the disadvantage of the latter methods is that they impose a dependency structure between labels, and hence learning will only be possible with dependencies that respect this structure.

Generally, infectious diseases share common spread risk factors; this can lead to interdependencies between them. For example, a recent paper [13] showed that the presence of enclosures or parks is a risk factor for spread of Crimean-Congo, Rift Valley Fever and Dengue. Similarly, temperature variations are a common risk factor for spread of Covid-19, Meningitis and Measles. Thus, resurgence of one of these diseases may increase the likelihood resurgence or emergence of others, particularly in the Senegalese context, where various propagation risk factors are present in health districts such as: winter season, wind and dust, promiscuity, humidity, temperature variations, gatherings in public places, etc.

The two predictive models we propose in this work are described below. Their advantage is that they take account of interdependencies between labels, without fixing any particular dependency structure. Moreover, they do not require the determination of the covariance structure of the labels and are well suited to qualitative data.

Multiple Binary Random Forest (MBRF): It consists of three steps:

- First, transform the multi-label classification problem into several binary classification problems where each label (disease) is considered separately as a binary variable.
- Second, to take account of interdependencies between diseases in each binary model, we add the other labels (or diseases) that are not predicted as predictors.
- Third, group all the individual outputs of the binary problems to obtain the predictions of all ten diseases.

The binary random forest uses the ranger function with Gini criterion. Hyper parameters such as: *mtry* and *min.node.size*, are selected via the *GridSearchCV* method, which provides the best values of these hyper parameters using a cross-validation procedure. *mtry* is the number of variables randomly selected for node splitting, and *min.node.size* is the minimum size of a terminal node.

Multi-output decision tree (MODT): It is an adaptation of the tree induction algorithm proposed by [5]. Instead of entropy, our algorithm utilizes an inertia criterion as the impurity measure. The information gain is then replaced by the inertia gain, which is calculated via Chi-square distance similarly to the calculation of inertia in the Multiple Correspondence Analysis technique. Suppose that the output variable Y has K modalities (here $K = 10$) and that the frequency of a modality j, $j = 1, \ldots, K$ in a node t is denoted p_jt. Then the inertia of node t is given by:

$$I(t) = \sum_{i \in t} \frac{1}{n_t} d^2(i, g_t), \tag{1}$$

where n_t is the node size, $g_t = (p_{1t}, \ldots, p_{Kt})$ is the center of gravity of node t, i represents an instance (or individual) in node t, and the Chi-square distance is

$$d^2(i, g_t) = \sum_{j=1}^{K} \frac{1}{p_{jt}} (y_{ij} - p_{jt})^2, \tag{2}$$

where $y_{ij} = 1$ if individual i takes modality j, and 0 elsewhere. The inertia $I(t)$ maybe interpreted as a measure of the homogeneity of node t.

2.3. Performance metrics

To assess the prediction quality of our two models, we use appropriate performance metrics, including: accuracy, kappa coefficient, precision, recall (sensitivity), and F1-score. Those metrics are calculated from the confusion matrix which comprises the following four categories:

- True Positive (TP): Number of instances predicted as positive for a disease, when they are actually positive for that disease.
- True Negative (TN): Number of instances predicted as negative for a disease, when they are actually negative for that disease.
- False Positive (FP): Number of instances predicted as positive for a disease, when they are actually negative for that disease.
- False Negative (FN): Number of instances predicted as negative for a disease, when they are actually positive for that disease.

Accuracy: Measures the probability of correct predictions among all predictions. It is calculated by:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}). \tag{3}$$

Kappa coefficient: Measures the agreement between the model predictions and the actual classes, taking into account the possibility of random agreement. It is calculated by:

$$\text{Kappa} = (P_o - P_e)/(1 - P_e), \tag{4}$$

where P_o is the observed accuracy and P_e is the accuracy expected by chance.

Precision: Measures the probability of true positives among all instances predicted as positive. It is calculated by:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}). \tag{5}$$

Recall/Sensitivity: Evaluates the model's ability to correctly identify all real positive instances. It is given by:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}). \tag{6}$$

F1-Score: Is the harmonic mean of Precision and Recall, providing a synthetic measure of the overall performance of the model that takes into account the balance between these two metrics. It is given by:

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})). \tag{7}$$

Since we deal with multi-class models, we must use an aggregation procedure to assess their performance. Here, we choose the micro-averaging procedure as the data are unbalanced, because Covid-19 is over-represented in the database. This procedure consists of gathering all true positives, false positives, true negatives and false negatives, respectively, over the different classes, and then applying the formula of the metric we wish to evaluate. For example, for the Precision, one has

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{K} \text{TP}_i}{\sum_{i=1}^{K} (\text{TP}_i + \text{FP}_i)}, \tag{8}$$

where TP_i is the number of true positives, FP_i the number of false positives in class i, and K is the number of classes.

3. Results and discussion

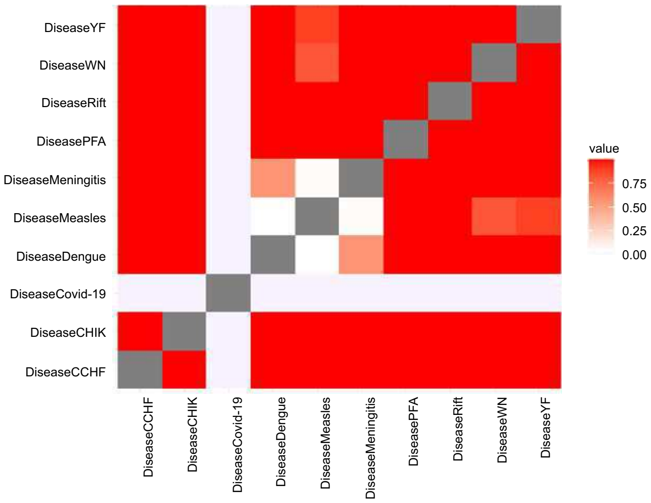

Figure 1. P-values matrix for the Chi-square test of associations between the ten infectious diseases.

We split the database into two parts: a training sample (90%) and a test sample (10%). To assess interdependencies between the ten diseases, we applied a Chi-square test to each couple of diseases; each disease is considered as binary variable. The resulting p-value matrix of the test (see Figure 1) indicates the statistical significance of the relationships between the diseases. A low p-value (usually $<= 0.05$) suggests a significant association, while a high p-value (> 0.05) indicates no association between the diseases.

The p-value matrix reveals that there are interdependencies between certain diseases, indicating that they are sharing common propagation risk factors. For example, Covid-19 is significantly related to almost all the other diseases. While Measles, Dengue and Meningitis are significantly associated with each other. This hypothesis of interdependency between the ten diseases justify our multi-output approach, and can lead to performance improvement. Table 2 displays performances of our two models and compare them with two standard models. The performance metrics were aggregated using micro-averaging procedure.

Table 2. Performance comparison.

Performance metrics	MODT	MCRF	MBRF	MLR
Accuracy	0.9945	0.9943	0.9999	0.9191
Kappa	0.8952	0.8927	0.8264	0.7316
Precision	0.8327	0.8197	0.9999	0.7938
Recall	0.8736	0.8722	0.9999	0.8504
F1-Score	0.8572	0.8451	0.9999	0.8211

Table 2 shows good predictive performance for all models, but our two proposed models MODT and MBRF seem to outperform the two others. It is also worth noting that the MBRF model outperforms the others on all metrics. This may be due to the fact that this model takes better account of the dependencies between diseases by considering, for each predicted disease, the other non-predicted diseases as predictors. To highlight the differences between the four models, we make a ranking via their performance, see Figure 2. By visualizing the performance metrics, we can observe again that the multiple binary random forest (MBRF) exhibits the best performance, followed by the multi-output decision tree (MODT) and the multi-class random forest (MCRF). The multinomial logistic regression (MLR) shows less performance.

The previous observations may lead us to validate our two proposed models, and use them to predict the resurgence of each of the ten diseases. Table 3 displays the resurgence probabilities of the ten diseases for both models, obtained by taking the proportion of positive predictions of each disease in the test sample.

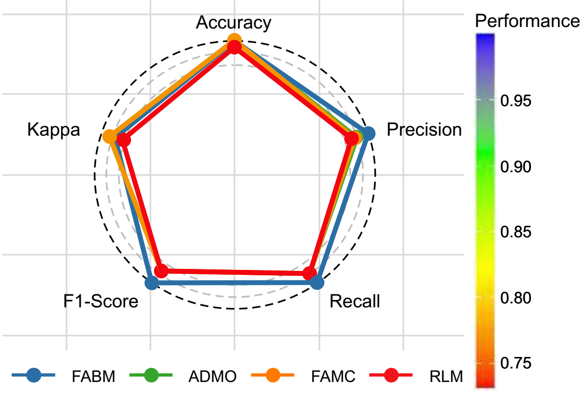

Figure 2. Model ranking.

From Table 3, we can say that without Covid-19, these infectious diseases have an average resurgence probability of 12.2% except for Poliomyelitis which recorded a lower resurgence probability of 2.39%. These results are in line with those of [14] who found that, out of 24 296 cases of fever, 11% were related to arboviruses/hemorrhagic fevers. In addition, [14] found that during the study period, Senegal has experienced a high incidence of infectious disease epidemics; some of which had already been eradicated (Poliomyelitis, Measles), while others are emerging due to climate change and environmental modification (Chikungunya, Rift Valley Fever, Crimean-Congo Hemorrhagic Fever, Covid-19). However, in the presence of Covid-19, we noted that the resurgence probability decreases to around 5% for all the other diseases. This could be explained by the fact that during the Covid-19 pandemic, surveillance was much more focused on the latter, thus relegating other existing pathologies to the background; see, e.g., [15,16]. Since Covid-19's instances are over-represented in the database, another explanation might be the fact that predictive models often favor the majority class.

Table 3. Probability of resurgence of the ten diseases obtained with MBRF and MODT models.

Infectious diseases	MBRF		MODT	
	With Covid-19	Without Covid-19	With Covid-19	Without Covid-19
CCHF	0.052426993	0.1232745	0.051785046	0.13573521
CHIK	0.049948422	0.12202076	0.050535339	0.12694722
Covid-19	0.603948285	XXXXXXX	0.606208411	XXXXXXX
Dengue	0.050957858	0.12186089	0.050591703	0.12589665
Measles	0.033162584	0.08962416	0.036245018	0.09044603
Meningitis	0.047739217	0.13515262	0.047580506	0.11302989
PFA	0.009798079	0.02329153	0.009227454	0.02392803
Rift	0.050830319	0.12578174	0.049016127	0.13505944
WN	0.050863152	0.12742476	0.048928176	0.13036992
YF	0.050492523	0.13199979	0.049882221	0.11858761

4. Conclusion

In this article we presented two new machine learning models to predict the resurgence of ten potentially epidemic infectious diseases under surveillance in Senegal. The first model MBRF (multiple binary random forest) treats separately each disease, but takes account of its interdependence with the others. The second model MODT (multi-output decision tree) introduces a specific impurity measure based on inertia, and allows to jointly predict all of ten diseases. These two models enable us to estimate the resurgence probability of each disease, which can be interpreted as the incidence rate of that disease, and thus allows to assess frequency and speed of the distribution of that disease over a given period.

One limitation of these models is that they do not take into account the temporal evolution of the data. As data were collected weekly, an interesting perspective would be to consider time series or LSTM (Long Short-Term Memory) models to improve predictions. Also, not all factors that are likely to influence the spread of these diseases have been taken into consideration; those might be: climate change, pathogens evolution, population movements (due to crises, wars, flooding, etc.), air traffic. Thus, a promising prospect would be to integrate more spread risk factors and explore deep learning or Bayesian approaches.

[1] Chen X., Moraga P. Forecasting dengue across Brazil with LSTM neural networks and SHAP-driven lagged climate and spatial effects. BMC Public Health. **25**, 973 (2025).

[2] Santangelo O. E., Gentile V., Pizzo S., Giordano D., Cedrone F. Machine Learning and Prediction of Infectious Diseases: A Systematic Review. Machine Learning and Knowledge Extraction. **5** (1), 175–198 (2023).

[3] Wang C., Zhou J., Huang H., Shen H. Classification Algorithms for Unbalanced High-Dimensional Data with Hyperbox Vertex Over-Sampling Iterative Support Vector Machine Approach. 2020 Chinese Control And Decision Conference (CCDC). 2294–2299 (2020).

[4] Evgeniou T., Pontil M. Regularized multi–task learning. KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 109–117 (2004).

[5] Linusson H. Multi-Output Random Forests. Dissertation. University of Bores/School of Business and IT (2013). `https://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-17167`.

[6] Glocker B., Pauly O., Konukoglu E., Criminisi A. Joint Classification-Regression Forests for Spatially Structured Multi-object Segmentation. Computer Vision – ECCV 2012. 870–881 (2012).

[7] Segal M. R. Tree-structured methods for longitudinal data. Journal of the American Statistical Association. **87** (418), 407–418 (1992).

[8] De'ath G. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology. **83** (4), 1105–1117 (2002).

[9] Segal M., Xiao Y. Multivariate random forests. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. **1** (1), 80–87 (2011).

[10] Zhang H. Classification Trees for Multiple Binary Responses. Journal of the American Statistical Association. **93** (441), 180–193 (1998).

[11] Read J., Pfahringer B., Holmes G., Frank E. Classifier Chains for Multi-label Classification. Machine Learning and Knowledge Discovery in Databases. 254–269 (2009).

[12] Read J., Martino L., Olmos P. M., Luengo D. Scalable multi-output label prediction: From classifier chains to classifier trellises. Pattern Recognition. **48** (6), 2096–2109 (2015).

[13] Ndao A., Seck C. T. Identification of propagation risk factors for ten infectious disesases under surveillance in Senegal. African Journal of Applied Statistics. **11** (1), 1535–1551 (2024).

[14] Barry A., Sagne S., Talla C., et al. Surveillance sentinelle des maladies 'a potentiel épidémique au Sénégal, Revue d'Epidémiologie et de Santé Publique. **71** (3), 101872 (2021).

[15] Dieng A., Diouf J. B. N., Ndiaye S. M. L. COVID-19 au Sénégal: réflexion d'un microbiologiste. Pan African Medical Journal. **35** (2), 31 (2020).

[16] Diouf I., Bousso A., Sonko I. Gestion de la pandémie COVID-19 au Sénégal [COVID-19 pandemic management in Senegal]. Médecine De Catastrophe – Urgences Collectives. **4** (3), 217–222 (2020).

Прогнозування повторного спалаху десяти інфекційних хвороб, що перебувають під наглядом у Сенегалі

Ндао А.[1], Сек К. Т.[1], Діоп Б.[2]

[1] *Університет Аліуна Діопа, а/с 30, Бамбей, Сенегал*
[2] *Управління профілактики, Міністерство охорони здоров'я, Дакар, Сенегал*

У цій роботі запропоновано дві прогностичні моделі для оцінки ймовірності повторного спалаху (резургенції) десяти інфекційних хвороб, що перебувають під епідеміологічним наглядом у Сенегалі. Перша модель — це множинний бінарний випадковий ліс (MBRF), який використовує функцію ranger із критерієм Джині та дозволяє окремо прогнозувати кожну з десяти хвороб, враховуючи їхні взаємозалежності. Друга модель — це мульти-вихідне дерево рішень (MODT), яке впроваджує критерій інерції (розрахований за допомогою відстані хі-квадрат) як міру чистоти вузла і дозволяє одночасно прогнозувати всі десять хвороб. Дані отримані з глобальної бази даних епідеміологічного нагляду Міністерства охорони здоров'я та містять інформацію про 68 698 випадків, що стосуються характеристик захворювань, районів, а також пацієнтів. Результати показали, що протягом періоду дослідження (січень 2018 р. — листопад 2022 р.) ці десять патологій зафіксували середню ймовірність повторного спалаху на рівні 12.2%, за винятком поліомієліту, показник якого був нижчим і становив 2.4%, та COVID-19, який продемонстрував досить високий рівень резургенції — близько 60%. Порівняно зі стандартними алгоритмами, такими як багатокласові випадкові ліси (MCRF) та мультиноміальна логістична регресія (MLR), дві запропоновані моделі забезпечили кращу ефективність. Наприклад, за показником F1-score отримано такі результати: MBRF (0.9999), MODT (0.8572), MCRF (0.8451), MLR (0.8211).

Ключові слова: *прогностичні моделі; багатоміткові моделі; ймовірність повторного спалаху; інфекційні захворювання.*