МАТЕМАТИЧНЕ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

MATHEMATICAL AND COMPUTER MODELING

UDC 004.94

MODELING OF THE SPREAD OF TUBERCULOSIS BY REGIONS IN UKRAINE

Boyko N. I. – PhD, Associate Professor, Associate Professor of the Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine.

Rabotiahov D. S. – Student, Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. Modelling the spread of tuberculosis in Ukraine is particularly relevant due to the increasing number of cases, especially in 2023.

Objective. The aim of this study is to solve modeling tasks by applying modern machine learning methods and data analysis to build predictive models of tuberculosis spread at the regional level.

Method. To model the spread of tuberculosis at the regional level in Ukraine, it is proposed to use several approaches, such as the SIR model, cellular automata, and Random Forest. Each of these methods has its unique advantages and can provide a more detailed understanding of the dynamics of disease spread. The SIR model (Susceptible-Infectious-Recovered) is a classical epidemiological model that describes the spread of infectious diseases in a population. The model assumes three groups of the population: *S* (Susceptible) – susceptible to infection; *I* (Infectious) – infected and capable of transmitting the infection; *R* (Recovered) – those who have recovered and gained immunity. Cellular automata are a discrete model that uses a grid of cells to simulate spatiotemporal processes. Each cell can be in different states (e.g., healthy, infected, immune) and change its state depending on the states of neighboring cells. Random Forest is a machine learning method that uses an ensemble of decision trees for classification or regression. This method can be applied to predict the spread of tuberculosis based on a large number of input parameters. Using these methods will allow for a deep analysis and comprehensive results regarding the spread of tuberculosis at the regional level in Ukraine. This, in turn, will facilitate the development of effective strategies to combat the disease and improve public health..

Results. The results of applying the Random Forest and SIR methods were described and analyzed in detail. For Random Forest, the metrics MSE and R^2 were evaluated, showing high prediction accuracy. In the case of the SIR algorithm, visual assessment of the results revealed insufficient accuracy due to model limitations. Comparing the chosen methods with other studies, a conclusion was made about the need to consider more complex algorithms to obtain more accurate results.

Conclusions. Based on the research results, it can be concluded that the Random Forest method is sufficiently effective for predicting vulnerable social groups and that the SIR algorithm is less effective for modeling the spread of tuberculosis. For further research development, it is recommended to consider more complex algorithms and account for additional factors influencing the spread of the disease. Moreover, to better understand further actions to combat the disease, it is advisable to simulate the spread of tuberculosis among the population of Ukraine.

KEYWORDS: method, metric, tuberculosis, Random Forest, Susceptible-Infectious-Recovered, modeling, algorithm.

ABBREVIATIONS

TB is a tuberculosis;
WHO is a World Health Organization;
DT is a Decision Tree;
SIR is a Susceptible-Infectious-Recovered;
MSE is a Mean Squared Error.

INTRODUCTION

Modeling the spread of tuberculosis in Ukraine at the regional level in 2024 is particularly relevant due to the increasing number of cases, especially in 2023. Tuberculosis remains one of the significant global problems of modern society, particularly in the context of the pandemic, and requires a comprehensive approach to its study and control [1, 3].

The scientific development of this problem includes various methods and approaches. Currently, statistical methods, epidemiological models, and machine learning methods are already being used to analyze and predict the spread of the disease. However, there are significant gaps in understanding the dynamics and specifics of tuberculosis spread at the regional level in Ukraine [5].

Object of the study: The social and demographic aspects of tuberculosis spread by gender and age.

Subject of the study: The methods and algorithms for studying the spread of tuberculosis in the regions of Ukraine.

DOI 10.15588/1607-3274-2024-4-4

The aim of this study is to address these issues by applying modern machine learning methods and data analysis to build predictive models of tuberculosis spread at the regional level. The conclusions drawn from this research can make a significant contribution to the development of effective strategies for disease control and prevention in Ukraine.

Tasks of the research:

– Data collection and preparation: Assess the available data on the number of tuberculosis cases in each region of Ukraine for 2024, considering age and gender distribution.

– Data analysis: Study the distribution of tuberculosis cases by region, age, and gender to identify possible dependencies and correlations.

– Develop a predictive model: Apply machine learning methods, particularly the Random Forest algorithm, to build a predictive model for tuberculosis spread at the regional level. Consider risk factors such as age and gender to identify the most vulnerable population groups.

– Evaluate model effectiveness: Analyze the results and evaluate the accuracy and reliability of the predictive model. Identify the most vulnerable population groups.

The advantages of modeling for such tasks lie in the ability to predict disease spread dynamics, identify the most at-risk population groups, and determine effective control strategies. Models allow for the consideration of various factors, such as demographic and socio-economic characteristics, which helps in understanding the complex relationships affecting disease spread. They can also be a useful tool for decision-making and developing preventive and treatment strategies.

The relevance of the topic reflects the importance and significance of the problem being studied and its alignment with contemporary scientific and practical needs. Modeling the spread of tuberculosis in Ukraine is particularly relevant due to the increasing number of cases, especially in 2023. The study is driven by the high incidence rate and potential threat to public health, which requires thorough analysis and effective management strategies [2].

Understanding the spread of tuberculosis and identifying the most vulnerable population groups is crucial for further control and prevention of this disease. Research in this area not only provides a scientific component but also has a direct practical impact on the health of citizens and the healthcare system of the country [6].

Moreover, modeling the spread of tuberculosis with the identification of the most vulnerable population groups by gender and age is significant for the further development of medical science and practice. Discovering new connections and factors influencing disease spread can contribute to improving tuberculosis diagnosis and treatment methods, as well as developing effective control and prevention programs.

The main focus of the research is on identifying the connections between various social and demographic factors and the spread of the disease, as well as determining the factors contributing to the risk of illness among different population groups. Thus, the object and subject of the research reflect the key aspects investigated within this work to address the problem of tuberculosis spread and improve public health.

1 PROBLEM STATEMENT

The relevance of this research lies in its ability to forecast disease spread dynamics, identify the most at-risk population groups, and determine effective control strategies. The models developed will consider various factors, such as demographic and socio-economic characteristics, to understand the complex relationships affecting disease spread. These models can also serve as a valuable tool for decision-making and developing preventive and treatment strategies.

Let:

– N be the total population of a region;

– $S(t)$ be the number of susceptible individuals at time t;

– $I(t)$ be the number of infectious individuals at time t;

– $R(t)$ be the number of recovered individuals at time t;

– β be the transmission rate;

– γ be the recovery rate.

The **SIR** model is described by the following set of differential equations [5, 7]:

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)I(t)}{N},$$

$$\frac{dI(t)}{dt} = \beta \frac{S(t)I(t)}{N} - \gamma I(t),$$

$$\frac{dR(t)}{dt} = \gamma I(t).$$

Objective Function for SIR Model: Minimize the error between the model predictions and the actual data [8, 9]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[I_{actual}(t_i) - I_{predicted}(t_i)]^2.$$

For the **Random Forest** model, let:

– X be the feature matrix (including age, gender, socio-economic factors, etc.),

– y be the target variable (number of tuberculosis cases).

The objective is to train the Random Forest model to minimize the mean squared error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2,$$

where $\hat{y}_i$ is the predicted value from the Random Forest model.

DOI 10.15588/1607-3274-2024-4-4

Using these methodologies, we aim to develop accurate predictive models to understand the spread of tuberculosis and identify the most vulnerable groups, thereby contributing to better disease control and prevention strategies.

2 LITERATURE REVIEW

To perform this research there will be used various methods aimed at analysing the socio-demographic aspects of the spread of tuberculosis and identifying the most vulnerable groups of the population. One of the main methods is the analysis of statistical data, which will provide objective results on the distribution of the disease among different categories of the population. Machine learning methods such as regression or clustering algorithms will also be used to identify complex relationships and patterns in the data. To analyse the geographical distribution of the disease, datasets on different regions over a long period of time will be used. In addition, epidemiological models will be used to predict the spread of tuberculosis and evaluate the effectiveness of control strategies. Such a comprehensive approach to research will allow us to gain a deeper understanding of the problem and identify the best ways to combat the disease.

Ilnytskyi H. I. (2021) [4, 10] in his research work investigated the epidemiological situation with tuberculosis in Ukraine. In addition to analysing statistical data and conducting sociological research, the author also developed a mathematical cellular SIS model to predict the spread of tuberculosis in Ukraine. The methods are based on a system of differential equations that describe the dynamics of tuberculosis spread in a population. The methods take into account factors such as TB incidence, TB mortality, infection rates and treatment effectiveness. This approach is not very suitable for the real world, as it assumes that an individual becomes infectious immediately after infection, which is not true, as the disease has an incubation period. The author also considered the SEIS model. The SEIS model considered here can be interpreted as an SIS model with an effective delay in the spread of the disease. It performed better under conditions close to those of the real world.

Researcher Rogovskyi V. O. (2023) [11, 18] developed a mathematical model for predicting the success of tuberculosis treatment. The model is based on a system of differential equations that describe the dynamics of the number of tuberculosis pathogens in the body. The model takes into account the following factors: the sensitivity of the pathogen to chemotherapeutic drugs, the patient's immune status, and the patient's compliance with the chemotherapy regimen. As analysed by the author, the model makes good predictions for short periods of time, as evidenced by the margin of error, but applying this model over longer periods of time can lead to an increase in the prediction error.

According to the Public Health Centre of the Ministry of Health of Ukraine [5, 12], 23,788 new cases of tuberculosis were registered in 2022, which is 60.1 per 100,000 people. Expectations for reducing TB morbidity and mortality in 2022: 10% reduction in morbidity and 5% reduction in mortality. Actual results: 8% reduction in incidence and 7% reduction in mortality.

The section "Creating a Bayesian Network of Risk Factors for COVID-19" of the thesis by Yaroslav Shevchenko [13, 15] describes two methods used to build a Bayesian network: the method of expert opinion (the author interviews 10 experts (infectious disease doctors) about their opinion on the impact of 12 risk factors on the likelihood of contracting COVID-19). Based on these opinions, a Bayesian network is built. Pros: simplicity and accessibility (the expert opinion method is easy to use and does not require special programming knowledge), flexibility (Bayesian networks can easily adapt to new data and information), visuality (Bayesian networks provide a convenient visual representation of the relationships between risk factors). Cons: subjectivity (the method of expert assessment can be subjective, depending on the opinion of experts), data requirements (the training algorithm of the Bayesian network requires a large amount of data), difficulty of interpretation (interpretation of the results of the Bayesian network can be difficult for people without special knowledge).

From the journal Chemistry, Ecology and Education, section "Mathematical modelling of the spread of viral infections in local urban ecosystems" [17, 20], one can learn about the process of spreading respiratory viral diseases, which are most often transmitted by airborne droplets. Infection by this route is most likely in crowded places. It is important to note that people start infecting others before they become ill. The following modelling techniques were used: a mathematical model (the authors developed a mathematical model to predict the spread of respiratory viral infections), Langevin dynamics (the model uses Langevin dynamics to model the movement of agents in the system), Levy flight modelling (the model allows for the sudden movement of infected agents over long distances). Advantages: adequacy (the model adequately reflects the main spatial and temporal components of the urban ecosystem), versatility (the model is universal and can be adjusted according to needs), effectiveness (the model can be used to predict epidemics and evaluate the effectiveness of prevention methods). Cons: Complexity (the model can be difficult to understand and use), need for the large amounts of data (data on the urban ecosystem and human behaviour are required to parameterise the model), limitations (the model cannot take into account all factors that affect the spread of infections).

Artificial intelligence can be a powerful tool for predicting the spread of TB and developing effective control strategies. An article titled "Prospects for the application of artificial intelligence to predict the spread of tuberculosis infection in the WHO European Region" [14, 19] only confirms this. In the study, the authors describe various AI methods that can be used to predict TB. TB spread models, that were mentioned in the paper:

– the classic SIR model:

– The model uses three states for agents: susceptible, infected, and recovered.

DOI 10.15588/1607-3274-2024-4-4

– The model takes into account factors such as the number of agents, speed of movement, probability of infection, duration of the disease, etc.

– Urban environment model:

– The authors propose to develop a model that takes into account the life, behaviour and interaction of people in the city.

– This model will be integrated with the infection spread model for more accurate forecasting.

– Advantages of the proposed approach:

– Better forecasting: The combination of models will allow for more accurate forecasting of the spread of the epidemic at the regional, national and global levels.

– Incorporating geospatial data: The use of geographic maps and the location of buildings will allow us to study the spatial spread of the epidemic.

– Speed of calculations: The model should be fast enough to run a large number of computer experiments.

– Parallel computing: The possibility of parallel computing will allow modelling the spread of the epidemic at the macro level of countries and the world.

– Disadvantages of the proposed approach:

– Complexity of development: Developing an urban environment model is a complex task.

– The model must take into account many factors such as age, immunity, building types, etc.

– Data requirements: Large amounts of data are required to develop and train the model.

Table 1 provides an analysis of the research on this topic, which is subject to careful scientific review in order to understand the direction of change.

Table 1 – Review of related papers

Title of the work (author)	Methodology	Pros of the methodology	Cons of the methodology
Ilnytskyi H. I. (2021)	Mathematical cellular model SIS/SEIS	Simple, describes the dynamics of TB spread well	Does not take into account the incubation period, not very suitable for the real world
Rogovskyi V. O. (2023)	Mathematical model of the dynamics of the number of TB pathogens	Allows to take into account susceptibility to chemotherapy, immune status, adherence to the regimen	Not very accurate over long periods of time
Public Health Centre of the Ministry of Health of Ukraine (2022)	Expectations and actual results in reducing morbidity and mortality from tuberculosis	A simple method of forecasting	Does not take into account the impact of various factors
Shevchenko Yaroslava (2020)	Bayesian network of risk factors for COVID-19	Simple, flexible, visual	Subjective, requires a lot of data, difficult to interpret

Summing up the results presented in Table 1, the analysis of the literature confirms the relevance of the problem of studying tuberculosis in Ukraine. Despite a certain decline in the incidence rate in recent years, the country's TB rate remains higher than in the European Union.

3 MATERIALS AND METHODS

In order to predict the spread of tuberculosis and to model it, we need a dataset that contains sufficient information to make the predictions accurate. That is why the dataset [17, 20] was chosen for the paper, which contains data on the incidence of tuberculosis in Ukraine for a long period from 2007–2022. The data is presented in tabular format and describes the following characteristics: the number of cases by date and region, age and gender, TB form and treatment outcome.

Although the dataset does not provide direct information on the causes of TB, it does allow for the study of risk factors such as age, place of residence, socioeconomic status, and comorbidities.

The data describe both men and women of all age groups, with detailed age categorisation.

In addition to information on the forms of TB, the dataset also contains data on treatment outcomes, which allows for an assessment of treatment effectiveness.

The existing dataset will be split 7:3 for training and testing, respectively.

For a better understanding of the approaches and to solve the tasks, let us consider the proposed methods listed in Table 1

The methodology of the mathematical cellular model SIS/SEIS described by Ilnytskyi G. I. (2021) has the advantages of ease of use and a good description of the dynamics of the disease spread, however, an important negative factor of the methodology is that it does not take into account the incubation period of the disease, which is why its use in this work is inappropriate [16].

Researcher Rogovsky V. O. (2023) proposed a mathematical model of the dynamics of the number of tuberculosis pathogens, which perfectly allows taking into account sensitivity to chemotherapy, immune status, and compliance with the regimen, but the lack of need to include these parameters in the analysis, along with the insufficient accuracy of predictions over long periods of time, make this methodology inappropriate for use in this work.

The data provided by the Public Health Centre of the Ministry of Health of Ukraine (2022) only shows comparative statistics on the expected and actual results of reducing the incidence and mortality from tuberculosis.

To address this issue, Yaroslav Shevchenko's thesis proposes a Bayesian network of risk factors for contracting COVID-19. The advantages of this method are the flexibility of the model and a good opportunity to visualise the results, but the model requires a large data set, which contradicts the above description of the data set, and the results of the algorithm are difficult to interpret.

DOI 10.15588/1607-3274-2024-4-4

The problem of the need for a large data set was also encountered by the methods of Langevin dynamics and Levy flight modelling described in the journal Chemistry, Ecology and Education (2023), although they had such advantages as the versatility of the algorithm and its efficiency.

The article "Prospects for the application of artificial intelligence to predict the spread of tuberculosis infection in the WHO European Region" presents the classic SIR model, an urban environment model that has such advantages as the inclusion of geospatial data, calculation speed, and the possibility of parallel computing. There are also disadvantages of the methodology, including the complexity of development and the need for a medium to large data set.

However, for a comparative analysis, the methods described in Table 1 should be considered, as well as additional tools that are best suited to the task at hand.

To address the issue of predicting the most vulnerable populations by age and gender in terms of TB incidence, decision trees can be used, namely Random Forest. Random Forest is a machine learning algorithm that uses a combination of DTs to improve accuracy. It works well with medium-sized datasets, because in our case, the selected dataset contains information for each of the 24 regions of Ukraine and the city of Kyiv in the time period from 2007–2022, which makes it possible to track the dynamics of the disease in each of the regions over an average period of 15 years.

Random Forest can help us to:

– Identify risk factors: Random Forest can help identify factors (gender, age, other) that influence the likelihood of getting TB.

– Classifying people into risk groups: Random Forest can classify people into risk groups based on their likelihood of getting the disease.

– Identify the most vulnerable groups by gender and age: Random Forest can help you identify the populations that are most at risk for TB.

The model will make predictions about the most vulnerable groups based on the number of cases, age groups, gender, and year. As a result of processing the data, the model will output the age group and gender of people who may be most affected by the disease.

The model works as follows:

– Random sampling: A set of random subsets is generated from the data.

– Training of the DTs: A DT is trained for each subset of the data.

– Aggregation: The forecasts from all the DTs are combined to produce the final forecast.

Here is the pseudo-code of the algorithm:

```
Algorithm Random Forest: pseudocode
1:  To generate c classifiers:
2:  for i = 1 to c do
3:  Randomly sample the training data D with
    replacement to produce Di;
4:  Create a root node, Ni, containing Di,
5:  Call BuildTree(Ni)
6:  end for
7:  BuildTree(N):
8:  if N contains instances of only one class
    then
9:  return
10: else
11: Randomly select x% of the possible split-
    ting features in N
12: Select the feature F with the highest in-
    formation gain to split on
13: Create f child nodes of N, N1,.... , Nf,
    where F has f possible values (F1, ...,
    Ff)
14: for i = 1 to f do
15:      Set the contents of Ni, to Di, where
    Di, is all instances in N that match Fi
16:   Call BuildTree(Ni)
17: end for
18: end if
```

Random Forest algorithm:

– You need to set the following parameters:

– Number of trees;

– Depth of trees.

– Next, you need to train the model, for this(for each tree):

– Select a random subset of the data;

– Train the DT on the subset of data.

– The next step is to make a prediction, for this(for each tree):

– Make a prediction for a new data instance;

– Combine the predictions from all the trees.

Also, for a better understanding of how the algorithm works, we can look at the flowchart of its operation in Fig. 1.

In Fig. 1 the following notation is used:

V_n: This is the entire dataset used to train the algorithm. It consists of n data points, where each data point is represented by a pair (x_i, y_i). Here, xi is the feature vector for the ith data point and y_i is the corresponding objective value.

v: This is the number of decision trees to be created in the algorithm.

V_1, V_2, ..., V_k: These are the random data sets used to train each decision tree. Each data set V_i consists of k data points that are randomly selected from V_n.

k: This is the number of features that are randomly selected for each decision tree.

A_1, A_2, ..., A_k: These are the predictions made by each decision tree for a new data instance X.

A: This is the final prediction made by the Random Forest algorithm. It is calculated as the average of the predictions made by all decision trees.

The R^2-measure (coefficient of determination) can be used to assess the accuracy, and the mean square error (the average of the squared differences between the predictions and the actual values), which will be referred to as MSE, can be used to assess the model error.

DOI 10.15588/1607-3274-2024-4-4

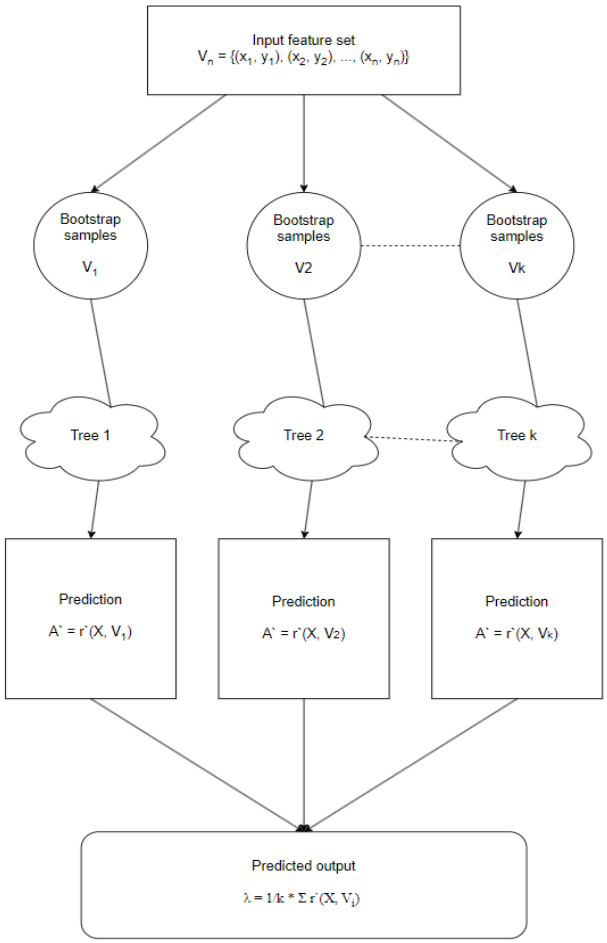

Figure 1 – Block diagram of the Random Forest algorithm

The next step, after identifying the most vulnerable part of the population, is to model the spread of TB in the regions of Ukraine. To perform this part of the task, it was decided to use the SIR model. The SIR model is a simple cellular automata model used to simulate the spread of infection. It divides people into three groups:

– Susceptible (*S*): People who can contract an infection.

– Infected (*I*): People who are infected with the infection and can transmit it to others.

– Recovered (*R*): People who have recovered from an infection and can no longer be infected.

The SIR model works as follows:

– Infected people transmit the infection to susceptible people with a certain probability.

– Susceptible people who are exposed to the infection become infected.

– Infected people eventually recover.

Fig. 2 shows an image of a compartmental diagram of the SIR model, where N, S_0, I_0, R_0 are the total population, the initial value of susceptible people , the initial value of infected people, and the initial value of people who can no longer get sick, respectively, β is a symbol is the symbol for the infection rate (describes the probability that a

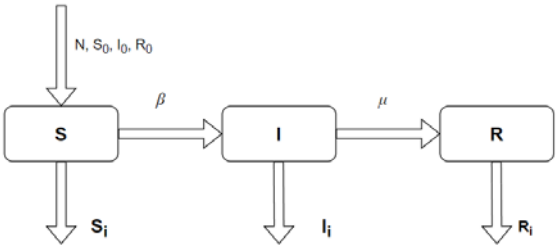

Figure 2 – Compartmental diagram of the SIR model

susceptible individual will become infected through contact with an infected individual), μ is the symbol for the mortality/cure rate (describes the probability that infected individuals will die of the disease or recover without the possibility of contracting the disease again).

The values of S_i, I_i, and R_i are calculated from the differential equations at each iteration of the simulation as follows, respectively [12, 15]:

$$S_i = \frac{-\beta}{N * S_{i-1} * I_{i-1}},$$

$$I_i = \frac{\beta}{N * S_{i-1} * I_{i-1} * (I - \mu)},$$

$$R_i = I * \mu.$$

Here is a pseudo-code of the algorithm:

```
Algorithm SIR: pseudocode
1:  function MAIN
2:  set t, fixed = (p^(fix), t), bounds = range for
    p^(cal)
3:  for run = 0 to Number_of runs parallel do
4:    call single run (bounds, fixed)
5:  end for
6:  read and assembles results from single runs
7:  end function
8:  function SINGLE_RUN(bounds, fixed)
9:  call differential_evolution (Objective func-
    tion, bounds, fixed)
10: save results of a single run
11: end function
12: function OBJECTIVE_FUNCTION(p, t)
13: s = SIRmodel (p)
14: objfun=O_y(s),t(p)
15: return objfun
16: end function
17: function SIRMODEL(p)
18: initialize S_nini
19: for n = n_ini + 1 to n_ini + N^(mod) - 1 do
20:   compute s_n from s_n-1 by (10)
21: end for
22: return s
23: end function
```

Before dividing the data into training and testing subsets, it is necessary to filter the data. To do this, it is necessary to delete the tables that will not be used in the analysis, for example “Block 4. Information about TB facilities”, “Block 5. TB treatment”, “Block 6. Bed capacity of TB facilities”, and other separate tables. Other tables should be combined by years and regions, and the data should be cleaned (filling in gaps if any, normalisation, standardisation, etc.). Examples of table blocks divided by the information they store are shown in Fig. 3.

DOI 10.15588/1607-3274-2024-4-4

Block 3. Tuberculosis / HIV infection	
Table 44	Incidence of tuberculosis in combination with AIDS (new cases + relapses)
Table 45	Registration of HIV-positive persons with tuberculosis
Table 46	Tuberculosis patients died from the disease caused by AIDS
Table 47	Prevalence of all forms of active tuberculosis in combination with the disease caused by HIV
Block 4. Availability of doctors in anti-tuberculosis institutions	
Table 48	Availability of phthisiologist doctors in the institutions of the Ministry of Health of Ukraine
Table 49	Medical positions in medical and preventive institutions of the Ministry of Health of Ukraine, 2021
Block 6. Bed fund of anti-tuberculosis institutions	
Table 69	The network of anti-tuberculosis healthcare institutions of the Ministry of Health of Ukraine
Table 70	Availability of hospital beds for tuberculosis patients in health care institutions of the Ministry of Health of Ukraine
Table 71	Indicators of the use of the bed fund of anti-tuberculosis health care institutions of the Ministry of Health of Ukraine, 2021
Table 72	Hospital and sanatorium care for tuberculosis patients by the territorial location of health care institutions of the Ministry of Health of Ukraine

Figure 3 – Examples of table blocks from the data set

– To perform the analysis, it is necessary to design 2 data sets. The first set, which will be used to predict the age and sex categories of people who are most vulnerable to further infection using the Random Forest algorithm, will consist of the following columns: AgeCategory, Year, Sex, TotalPopulation, Infected, Dead. An example from this dataset is shown in Fig. 4.

AgeCategory	Sex	Year	TotalPopulation	Infected	Dead
0–1	M	2007	46650000	15	2
1–4	M	2007	46650000	113	0
5–9	M	2007	46650000	107	1
10–14	M	2007	46650000	99	1
15–17	M	2007	46650000	300	1
18–24	M	2007	46650000	2445	162
25–34	M	2007	46650000	6195	1365
35–44	M	2007	46650000	6310	2405
45–54	M	2007	46650000	5958	2890
55–64	M	2007	46650000	2831	1195
65–100	M	2007	46650000	1895	586
0–1	F	2007	46650000	7	1
1–4	F	2007	46650000	83	3
5–9	F	2007	46650000	90	0
10–14	F	2007	46650000	107	1
15–17	F	2007	46650000	303	1
18–24	F	2007	46650000	1707	91
25–34	F	2007	46650000	2821	464
35–44	F	2007	46650000	2118	512
45–54	F	2007	46650000	1482	449
55–64	F	2007	46650000	806	171
65–100	F	2007	46650000	1303	205
65–100					

Figure 4 – An example of a designed data set for the Random Forest algorithm

Table 2 – Description of data set elements for the Random Forest algorithm

Feature name	Description
AgeCategory	Age categories represented in the original dataset from 0 to 100 years
Sex	Sex
Year	The year, used for reference
TotalPopulation	Country population
Infected	Amount of infected people
Dead	Amount of dead people

Table 2 presents a description of the features that make up the data set for the Random Forest algorithm.

For the dataset that will be used by the SIR algorithm to model the spread of the disease, we will design a dataset consisting of the following columns: Year, InfectedTB, DeadTB, Region, TotalPopulation, RecoveredTB. An example of the generated data is shown in Fig. 5.

Year	InfectedTB	DeadTB	Region	Total-Population	RecoveredTB
2022	405	93	Volyn	1018628	150
2022	2450	340	Dnipropetrovsk	3093176	1013
2022	133	13	Donetsk	1883713	58
2022	491	98	Zhytomyr	1179801	189
2022	880	136	Zakarpattia	1241643	358
2022	595	161	Zaporizhzhia	1637673	209
2022	378	44	Ivano-Frankivsk	1349096	161
2022	857	129	Kyiv	1789300	350
2022	607	100	Kirovohrad	897297	244
2022	13	78	Luhansk	666801	–32
2022	846	195	Lviv	2459763	313
2022	952	85	Mykolaiv	1091106	417
2022	2927	249	Odessa	2340332	1286
2022	700	70	Poltava	1344445	303
2022	530	67	Rivne	1140724	223
2022	271	88	Sumy	1033580	88
2022	280	29	Ternopil	1018462	121
2022	965	181	Kharkiv	2583325	377
2022	476	100	Kherson	1000166	181
2022	441	50	Khmelnytskyi	1225666	188
2022	419	95	Cherkasy	1157115	156
2022	406	42	Chernivtsi	887392	175
2022	469	78	Chernihiv	950773	188

Figure 5 – Example of a designed dataset for the SIR model

Table 3 – Description of data set features for the SIR model algorithm

Feature name	Description
Year	Year, used for reference
Region	Region of Ukraine (not included city of Kyiv)
RecoveredTB	Amount of recovered people

Table 3 provides a description of the features included in the dataset for the SIR algorithm.

DOI 10.15588/1607-3274-2024-4-4

4 EXPERIMENTS

Conducting experiments for the topic "Modelling the spread of tuberculosis in the regions of Ukraine for 2024 with the identification of the most vulnerable groups by gender and age" is of great importance, as it allows us to understand how accurately the selected models or algorithms have worked, namely, to understand how accurately the Random Forest model predicts the number of infections in the upcoming years, which we can use to make conclusions about the most vulnerable part of the population. In addition, the experiments make it possible to visually see the modelled dynamics of the spread of the disease in each of the country's regions using the SIR cellular automata model.

The software implementation was done using Python, a high-level general purpose programming language with a simple and readable syntax. It has a large number of libraries for a variety of tasks, making it a very powerful tool for solving problems in research, data analysis, and machine learning. Python's advantages include ease of learning, broad community support, and ease of use. In particular, the following libraries and tools were used to develop the application:

– Pandas is a data processing and analysis library that provides data structures and functions for working with them. It allows you to easily perform operations on large data sets, such as reading, writing, filtering, and aggregating data.

– NumPy is a library for scientific computing in Python. It provides support for arrays and mathematical functions, allowing you to easily perform calculations on numerical data.

– Scikit-learn is a machine learning library for Python. It contains implementations of many machine learning algorithms, such as classification, regression, clustering, and others, as well as tools for estimating and fitting model parameters.

– Seaborn is a Python data visualisation library based on the Matplotlib library. It provides high-level functions for creating attractive and informative graphs and charts.

– Matplotlib is a graphing and data visualisation library in Python. It allows you to create various types of graphs, such as line, pie, and bar charts, and customise their appearance.

– Scipy is a Python library for scientific and technical computing. It contains implementations of many algorithms for numerical computation, optimisation, signal processing, and other functions for working with scientific data.

The dataset for the Random Forest algorithm described in the previous section, an example of which can be seen in Fig. 6, is stored in .csv format, so you need to write a function that reads data from the file and returns it in the pandas Data frame format:

```
Function readDataFrame(file path):
1: initialize an empty DataFrame called df_rf
2: read the contents of the file using
   "pd.read_csv()"
3: add the contents of the read file to df_rf
4: return df_rf
```

№	Age-Category	Sex	Year	TotalPopu-lation	In-fected	Dead
0	0–1	M	2007	46650000	15	2
1	1–4	M	2007	46650000	113	0
2	5–9	M	2007	46650000	107	1
3	10–14	M	2007	46650000	99	1
4	15–17	M	2007	46650000	300	1
…	…	…	…	…	…	…
347	25–34	F	2022	41167000	753	63
348	35–44	F	2022	41167000	1119	134
349	45–54	F	2022	41167000	952	121
350	55–64	F	2022	41167000	644	79
351	65–100	F	2022	41167000	787	77

Figure 6 – Data set in the pandas DataFrame format

Fig. 7 shows an example of data from a loaded dataset. After reading, you need to check the data types that are stored in the set. The result of the check can be seen in Fig. 8.

For the model to work, you must first clean the data, check for spaces, and convert it to the required types, after which the data set looks like this:

AgeCategory	object
Sex	object
Year	int64
TotalPopulation	object
Infected	object
Dead	int64
dtype	object

Figure 7 – Data types of the downloaded dataset

№	AgeCa tegory	Sex	Year	TotalPopul ation	Infected	Dead
0	0–1	M	2007	46650000	15	2
1	1–4	M	2007	46650000	113	0
2	5–9	M	2007	46650000	107	1
3	10–14	M	2007	46650000	99	1
4	15–17	M	2007	46650000	300	1
…	…	…	…	…	…	…
347	25–34	F	2022	41167000	753	63
348	35–44	F	2022	41167000	1119	134
349	45–54	F	2022	41167000	952	121
350	55–64	F	2022	41167000	644	79
351	65–100	F	2022	41167000	787	77

Figure 8 – Data after cleaning

Fig. 9 shows an example of data after data cleaning operations.

AgeCategory	object
Sex	object
Year	int64
TotalPopulation	int64
Infected	int64
Dead	int64
dtype	object

Figure 9 – Data types after cleaning the dataset

DOI 10.15588/1607-3274-2024-4-4

Fig. 10 shows the types of data from the dataset for the Random Forest algorithm after data cleaning operations.

AgeCategory	0
Sex	0
Year	0
TotalPopulation	0
Infected	0
Dead	0
dtype	int64

Figure 10 – Checking for gaps in the data set

As can be seen from Fig. 10, the data in the dataset has no gaps, so we can proceed to normalise the required data columns, namely TotalPopulation, Infected and Dead, and perform label encoding for the Sex and AgeCategory columns. Normalisation will be performed using the following function:

```
   Function min_max_normalize(dataset column):
1: We look for the minimum value in the column
   and set it in the col_min variable
2: We look for the maximum value in the column
   and set it in the col_max variable
3: We calculate the updated value of the column
   elements according to the formula (column
   value - col_min) / (col_max - col_min)
4: We return the column with the updated values
```

The label encoding operation will be performed using the LabelEncoder function, which we import from the sklearn.preprocessing library.

After performing the operations, mentioned above, the new view of the dataset for the Random Forest algorithm can be seen in Fig. 11.

№	AgeCategory	Sex	Year	TotalPopulation	Infected	Dead
0	0	1	2007	1.0	0.00146	0.00068
1	1	1	2007	1.0	0.01574	0.00000
2	8	1	2007	1.0	0.01487	0.00034
3	2	1	2007	1.0	0.0137	0.00034
4	3	1	2007	1.0	0.043	0.00034
…	…	…	…	…	…	…
347	5	0	2022	0.0	0.109	0.02152
348	6	0	2022	0.0	0.162	0.04578
349	7	0	2022	0.0	0.13805	0.04134
350	9	0	2022	0.0	0.0932	0.027
351	10	0	2022	0.0	0.114	0.02631

Figure 11 – View of the dataset after its preparation

To better understand the relationship between the data in the dataset, let's illustrate the correlation table shown in Fig. 12.

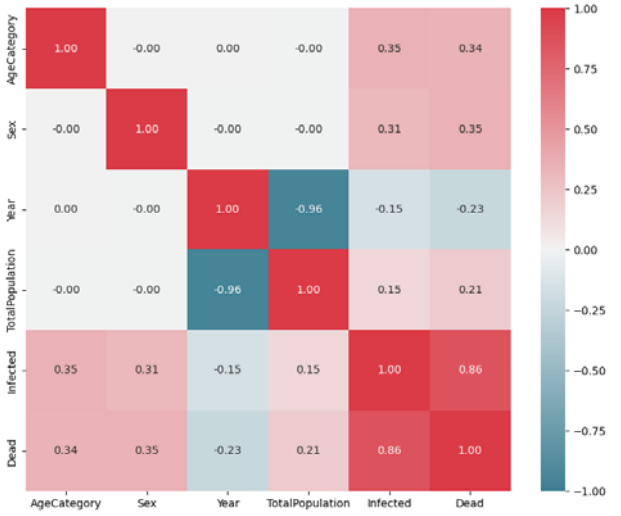

Figure 12 – Correlation table of the dataset

Cells in the correlation table with values from 0.7 to 1.0 will be considered strong correlations, while the medium correlation will be considered to be cells with values between 0.3 and 0.69. As we can see from Fig. 12, the features Infected and Dead, AgeCategory and Infected, Sex and Infected have an average correlation, while the features TotalPopulation and Year have a strong inverse correlation.

Before starting the algorithm, it is necessary to select the target variable, which will be the Infected feature, and all other features will be used for prediction. Also, it is necessary to divide the data set into training and testing, which will be done in the ratio of 7:3.

Finally, we can move on to the model that will be used for prediction. Given that the target feature is a continuous value, we used the RandomForestRegressor algorithm taken from the Python library sklearn.ensemble, where the number of trees to be used is set to 100 as a parameter.

Once the model is trained, we run the test, and evaluate the model's performance, using the metrics of root mean square error and R^2 estimation. The following results are obtained:

```
Mean Squared Error (MSE): 0.0028
R-squared: 0.92
```

As you can see, the value of the mean squared error is quite small, while the value of the R^2 estimate is quite high, with the highest possible value of this estimate being 1.

We also test the model by checking which population group, by age and gender, is most vulnerable to infection. We have the following results:

```
Based on the model, the group with the
highest predicted number of deaths in the
year 2024 is:
Age Category: 35-44
Sex: M
```


DOI 10.15588/1607-3274-2024-4-4

Based on these results, we can conclude that men aged 35 to 44 will have the highest number of infections in the next year. Also, to better understand the situation of infection for other categories of sex and age, we visualise the graph of the predicted spread of the disease among them, which is shown in Fig. 13.

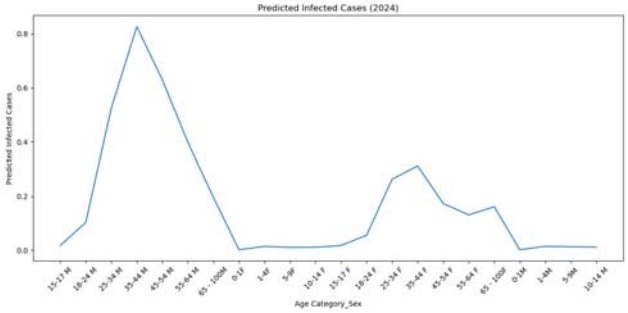

Figure 13 – Statistics of subsequent diseases for all categories of gender and age

From Fig. 13, we can see that the age group of men aged 35–44 is the most susceptible to infection, while the second place is occupied by the group of women of the same age category. The values for other age groups for both sexes were also illustrated in Fig. 13.

To model the spread of the disease in each of the 24 regions of Ukraine and in the city of Kyiv, it is necessary to use the data set, an example of which can be seen in Fig. 14. From the example of this data set, we have information on the number of infected, the number of recovered, the number of deaths, the year, the name of the region, and the population of the respective region.

Let's load the data set into the program in the pandas DataFrame format using the *readDataFrame* function.

№	Year	Infect-edTB	Dead TB	Region	Total-Popula-tion	Recov-eredTB
411	2022	441	50	Khmel-nytskyi	1225666	188
412	2022	419	95	Cherkasy	1157115	156
413	2022	406	42	Chernivtsi	887392	175
414	2022	469	78	Chernihiv	950773	188
415	2022	644	102	KyivCity	2910994	261

Figure 14 – Example of a data set in the DataFrame format

Fig. 14 shows an example of data from a dataset for the SIR algorithm after conversion to the DataFrame format. After reading, it is necessary to check the data types stored in the set, which results in the following result:

Year	int64
InfectedTB	object
DeadTB	int64
Region	object
TotalPopulation	object
RecoveredTB	int64
dtype	object

Figure 15 – Data types of the loaded dataset before cleaning

From Fig. 15, we can see that some numeric features that should represent numeric values do not have the required data type set, for which it is necessary to perform preliminary data cleaning and convert the required features to the required data types, and then we have:

Year	int64
InfectedTB	int64
DeadTB	int64
Region	string
TotalPopulation	int64
RecoveredTB	int64
dtype	object

Figure 16 – Data types of the loaded dataset after cleaning

Fig. 16 shows a description of the data types for the features of the dataset for the SIR algorithm after the preparatory operations. Also, let's check the data for completeness, that is, whether it has any empty fields. The result of this check is shown in Fig. 17.

Year	0
InfectedTB	0
DeadTB	0
Region	0
TotalPopulation	0
RecoveredTB	0
dtype	int64

Figure 17 – Checking the dataset for completeness

As we can see from Fig. 17, the data has no voids. Also, to better understand the relationships between the features in the dataset, let's illustrate the correlation table shown in Fig. 18.

Figure 18 – Correlation table for a data set

Describing the correlation table for the Random Forest dataset shown in Fig. 18, we defined the terms strong and medium correlation, so that the features InfectedTB and RecoveredTB, DeadTB and RecoveredTB, InfectedTB and DeadTB have a strong correlation, while almost all features at the intersection with the feature Year have a strong inverse correlation.

DOI 10.15588/1607-3274-2024-4-4

Before starting the SIR algorithm, it is necessary to describe the initial values of the parameters *S*, *I*, *R*, which stand for the number of susceptible people, the number of infected people, and the number of people who can no longer get sick. In the latter category, we included both people who have recovered and people who have already died from the disease. Among the parameters, there is also one that determines the number of days during which the disease spread will be modelled. In our case, it is set to 700 days. Also, for the algorithm to work, it is necessary to set the parameters β and μ, which act as coefficients in calculating the number of people moving from state *S* to state *I* (transmission coefficient or the rate at which susceptible people become infected when they come into contact with infected people. It indicates the rate at which the disease spreads in the population), and from state *I* to state *R* (the rate of recovery or the rate at which infected individuals recover from the disease and become immune), respectively. Thus, for the above parameters, which are necessary for the algorithm to work, we have set the following values:

S = TotalPopulation – InfectedTB – RecoveredTB – DeadTB;

I = InfectedTB;

R = number of recovered TB cases + number of dead TB cases;

β = 4/10 (which means that every 10 days 4 people become infected);

μ = 1/10 (which means that every 10 days one person recovers);

Finally, when the algorithm is finished, we have the following results of modelling the number of eligible, infected, recovered or dead people for some regions in Fig. 19–21:

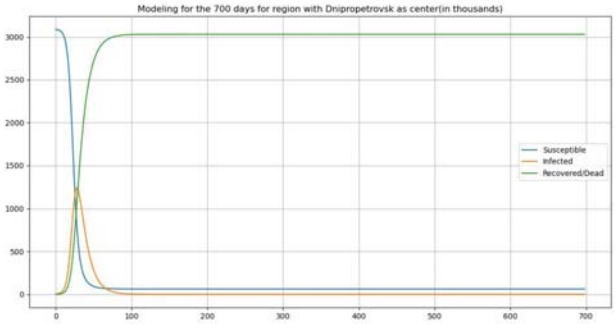

Figure 19 – Modelling results for Dnipro region

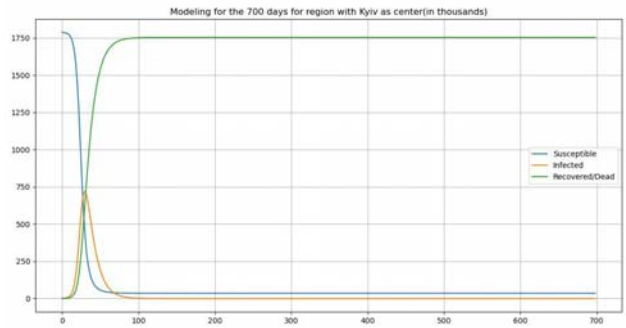

Figure 20 – Modelling results for Kyiv region

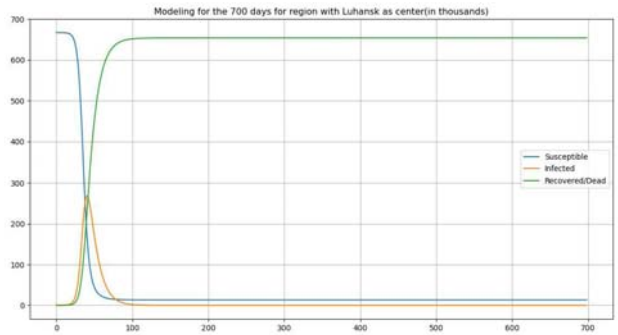

Figure 21 – Modelling results for Luhansk region

To consider examples of modelling results shown in Fig. 19–21, we have selected 3 regions of Ukraine that differ in population size, namely Dnipro region(population of almost 3 million people), Kyiv region(population of almost 1.8 million people) and Luhansk region(population of almost 0.7 million people). As we can see, regardless of the initial population (or people eligible for the disease), the largest increase in infected persons occurs in the first 100 days of the simulation, after which the number of people who can be infected and the number of infected persons drop to 0, moving to the recovered/dead state, the number of which is close to the initial number of people who can be infected.

To better understand the state of the disease in each of the oblasts in comparison to other oblasts, let us illustrate the maximum percentage of infected people in each of the oblasts for the entire modelling period relative to the total population in the oblast in Fig. 22.

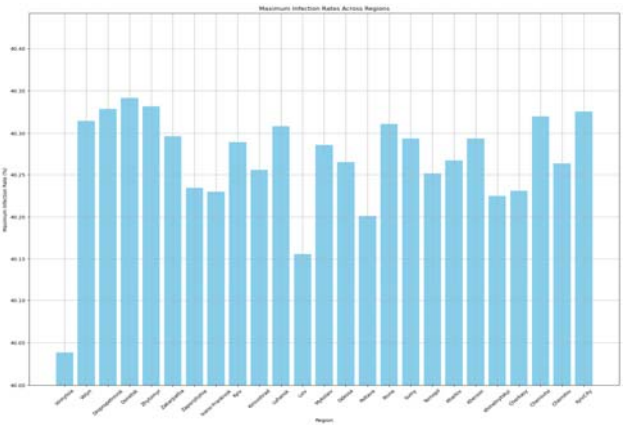

Figure 22 – Comparative histogram of the maximum number of infected people in relation to the total population of the region

From Fig. 22 shows that the maximum number of infected individuals relative to the total population of the region was approximately the same for the entire modelling period, fluctuating around 40%. The region with the lowest value of the maximum number of infected individuals is Vinnytsia, and the highest is Donetsk.

DOI 10.15588/1607-3274-2024-4-4

5 RESULTS

Chapter 4 described the algorithms and demonstrated their results. It is important not only to evaluate the results visually, but also with the help of metrics. As we can see from Section 4.2, the MSE and R^2 metrics were evaluated for the Random Forest algorithm. Under ideal conditions, the value of the MSE metric should be close to 0, while the value of the R^2 metric should be close to one.

Table 3 – Metrics values for the results of the Random Forest algorithm

Metric	Volue
MSE	0.0028
R^2	0.9204

As we can see from Table 3, the actual values of the metrics are quite close to the values of the metrics under ideal conditions, and the prediction provided by this algorithm is quite accurate, as can be seen in the statistical analysis of the data set.

The operation of the SIR algorithm was also described in section 4.3, after which, by calculating the results separately for each of the regions and visualising the maximum number of infected individuals relative to the total population of the region for the entire modelling period in Fig. 4.17, we were able to verify the lack of accuracy of the modelling of the disease spread, since the average value for all regions in Fig. 25 reaches 39%, which cannot be true in reality. Such inaccuracies in the results can be explained by the fact that risk factors for the disease, the possibility of moving from a state of recovery to a state of infection, and some others are not taken into account. Also, the parameters that are taken into account when transitioning between the states of healthy-infected and infected-recovered/dead, are an important part in this algorithm. These parameters need to be better tuned, which requires a larger scientific base.

In general, comparing the selected algorithms with those chosen by other researchers in related topics, we can clearly see that the selected models are architecturally the simplest in their field, and of course, their simplicity takes away from their high accuracy of results.

For further analysis of this topic, it is worth choosing more structurally complex algorithms, the accuracy of which will be higher, namely recurrent neural networks, or forecasting using time series.

6 DISCUSSION

Conducted research aimed at studying the spread, modeling of tuberculosis and forecasting the most vulnerable social groups of the population. During the analysis of the literature, the relevance of the problem was revealed and methods for modeling the disease were chosen.

A comparative analysis of other methods used in previous studies was conducted to determine the most effective methods in this context. The use of the Random Forest method has demonstrated sufficient effectiveness in predicting vulnerable social groups of the population in the context of the spread of tuberculosis.

The SIR algorithm proved to be less effective in modeling the spread of the disease due to its shortcomings identified during the study. It is necessary to consider more complex algorithms to obtain more accurate results in predicting the spread of tuberculosis.

For the further development of the research, it is recommended to take into account additional factors affecting the spread of the disease.

It is recommended to carry out a simulation of the spread of tuberculosis among the population of Ukraine for a better understanding of further actions in the fight against this disease.

CONCLUSIONS

Comparing the selected algorithms with other studies in related topics, one can understand that they are architecturally the simplest in their field.

The simplicity of these models takes away the ability to provide high accuracy results, but for further analysis of the topic, it is recommended to consider more architecturally complex algorithms, such as recursive neural networks or time series forecasting, which can provide even greater accuracy of predictions.

The study compared in detail the effectiveness of two different modeling approaches – machine learning (Random Forest) and epidemiological model (SIR). The Random Forest method was found to provide higher prediction accuracy, which is important for further research and practical work in the field of healthcare. A preliminary data analysis was carried out with the selection of the most suitable sets for the study, which increases the relevance and accuracy of the results obtained.

Practical significance of the results: Conclusions about the effectiveness of the Random Forest method can be used to create programs for predicting and controlling the spread of tuberculosis, which is an urgent task for the health care system.

Recommendations for further research: Based on the obtained results, recommendations have been developed regarding the use of more complex algorithms and consideration of additional factors that may affect the accuracy of models, which contributes to the development of a scientific approach to the study of the spread of diseases.

These aspects ensure the **scientific novelty** of the work and emphasize its significance in the context of tuberculosis research and forecasting of vulnerable social groups.

Prospects for further research are to study the proposed algorithms for a wide class of practical problems.

ACKNOWLEDGEMENTS

The study was created within research topic “Methods and means of artificial intelligence to prevent the spread of tuberculosis in wartime” (№0124U000660), which is carried out at the Department of Artificial Intelligence

DOI 10.15588/1607-3274-2024-4-4

Systems of the Institute of Computer Sciences and Information of technologies of the National University "Lviv Polytechnic".

REFERENCES

1. Duko B., Bedaso A., Ayano G. The prevalence of depression among patients with tuberculosis: a systematic review and meta-analysis, *Ann Gen Psychiatry*, 2020, Vol. 19(1), P. 1. Mode of access: https://doi.org/10.1371/journal.pone.0227472.
2. [Ruiz-Grosso P., Cachay R., De La Flor A. et al.] Association between tuberculosis and depression on negative outcomes of tuberculosis treatment: a systematic review and meta-analysis, *PLoS ONE*, 2020, Vol. 15(1), P. 1. Mode of access: https://doi.org/10.1371/journal.pone.0227472.
3. Cohen A., Mathiasen V. D., Schön T. et al. The global prevalence of latent tuberculosis: a systematic review and meta-analysis, *Eur Respir J*, 2019, Vol. 54(3), P. 1. Mode of access: https://doi.org/10.1183/13993003.00655-2019.
4. Kiazyk S., Ball T. Tuberculosis (TB): latent tuberculosis infection: an overview, *Canada Commun Dis Rep.*, 2017, Vol. 43(3–4), P. 62. Mode of access: https://doi.org/10.14745/ccdr.v43i34a01.
5. World Health Organization. Annual Report of Tuberculosis. Annual Global TB Report of WHO [Electronic resource]. Access mode: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 (access date: 11.05.2024). Title from the screen.
6. Abascal E., Pérez-Lago L., Martínez-Lirola M. et al. Whole genome sequencing-based analysis of tuberculosis (TB) in migrants: Rapid tools for crossborder surveillance and to distinguish between recent transmission in the host country and new importations, *Eurosurveillance,* 2018, Vol. 24(4), P. 1800005. Mode of access: https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800005.
7. De Beer J. L., Ködmön C., Van der Werf M. J. et al. Molecular surveillance of multi- and extensively drug-resistant tuberculosis transmission in the European Union from 2003 to 2011, *Eurosurveillance,* 2014, Vol. 19(11), P. 20742. Mode of access: https://doi.org/10.2807/1560-7917.ES2014.19.11.20742.
8. Dohál M., Dvořáková V., Šperková M. et al. Whole genome sequencing of multidrug-resistant Mycobacterium tuberculosis isolates collected in the Czech Republic, 2005–2020, *Sci Rep.*, 2022, Vol. 12(1), pp. 1–10. Mode of access: https://doi.org/10.1038/s41598-022-11287-5.
9. Dohál M., Dvořáková V., Šperková M. et al. Antituberculosis drug resistance in Slovakia, 2018–2019: the first whole-genome epidemiological study, *J Clin Tuberc Other Mycobact Dis*, 2022, Vol. 26, P. 100292. Mode of access: https://doi.org/10.1016/j.jctube.2021.100292.
10. Pavlenko E., Barbova A., Hovhannesyan A. et al. Alarming levels of multidrug-resistant tuberculosis in Ukraine: results from the first national survey, *Int J Tuberc Lung Dis,* 2018, Vol. 22(2), pp. 197–205. Mode of access: https://doi.org/10.5588/ijtld.17.0254.
11. Vyklyuk Ya. Nevinskyi D., Boyko N. GeoCity – a New Dynamic-Spatial Model of Urban Ecosystem, *J. Geogr. Inst. Cvijic*, 2023, Vol. 73(2), pp. 187–203. Mode of access: https://doi.org/10.2298/IJGI2302187V.
12. Zignol M., Cabibbe A. M., Dean A. S. et al. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study, *Lancet Infect Dis,* 2018, Vol. 18(6), pp. 675–683. Mode of access: https://doi.org/10.1016/S1473-3099(18)30073-2.
13. Paul L. K., Nchasi G., Bulimbe D. B. et al. Public health concerns about Tuberculosis caused by Russia/Ukraine conflict, *Health Sci Rep*, 2023, Vol.6(4), P. 1. Mode of access: https://doi.org/10.1002/hsr2.1218.
14. Aldridge R. W., Zenner D., White P. J. et al. Tuberculosis in migrants moving from high-incidence to low-incidence countries: a population-based cohort study of 519,955 migrants screened before entry to England, Wales, and Northern Ireland, *Lancet,* 2016, Vol. 388(10059), P. 2510. Mode of access: https://doi.org/10.1016/S0140-6736(16)31008-X.
15. Merker M., Blin C., Mona S. et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage, *Nat Genet,* 2015, Vol. 47(3), pp. 242–249. Mode of access: https://doi.org/10.1038/ng.3195.
16. Daum L. T., Konstantynovska O. S., Solodiankin O. S. et al. Next-generation sequencing for characterizing drug resistance-conferring mycobacterium tuberculosis genes from clinical isolates in the Ukraine, *J Clin Microbiol,* 2018, Vol. 56(6), P. 1. Mode of access: https://doi.org/10.1128/JCM.00009-18.
17. Huang C. C., Chu A. L., Becerra M. C. et al. Mycobacterium tuberculosis Beijing lineage and risk for tuberculosis in child household contacts, *Peru. Emerg Infect Dis,* 2020, Vol. 26(3), P. 568. Mode of access: https://doi.org/10.3201/eid2603.191314.
18. Vyklyuk Y., Semianiv I., Nevinskyi D. et al. Applying geospatial multi-agent system to model various aspects of tuberculosis transmission, *New Microbes and New Infections,* 2024, Vol. 59, P. 101417. Mode of access: https://doi.org/10.1016/j.nmni.2024.101417.
19. Stucki D., Ballif M., Egger M. et al. Standard genotyping overestimates transmission of mycobacterium tuberculosis among immigrants in a low-incidence country, *J Clin Microbiol*, 2016, Vol. 54(7), pp. 1862–70. Mode of access: https://doi.org/10.1128/JCM.00126-16.
20. Jackson S., Kabir Z., Comiskey C. Effects of migration on tuberculosis epidemiological indicators in low and medium tuberculosis incidence countries: a systematic review, *J Clin Tuberc Other Mycobact Dis,* 2021, Vol. 23, pp. 2405–5794. Mode of access: https://doi.org/10.1016/j.jctube.2021.100225.

Received 20.08.2024.
Accepted 24.10.2024.

DOI 10.15588/1607-3274-2024-4-4

УДК 004.94

МОДЕЛЮВАННЯ ПОШИРЕННЯ ТУБЕРКУЛЬОЗУ ЗА РЕГІОНАМИ В УКРАЇНІ

Бойко Н. І. – канд. економ. наук, доцент, доцент кафедри Систем штучного інтелекту, Національний університет «Львівська політехніка», Львів, Україна.

Работягов Д. С. – студент кафедри Системи штучного інтелекту, Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Моделювання поширення туберкульозу на території України є особливо актуальним у зв'язку зі зростанням числа випадків захворювання, зокрема у 2023 році.

Мета роботи є вирішення задач моделювання шляхом застосування сучасних методів машинного навчання та аналізу даних для побудови прогностичних моделей поширення туберкульозу на регіональному рівні.

Метод. Для моделювання поширення туберкульозу на регіональному рівні в Україні пропонується використовувати кілька підходів, таких як SIR модель, клітинні автомати та Random Forest. Кожен з цих методів має свої унікальні переваги та може забезпечити детальніше розуміння динаміки поширення захворювання. SIR модель (Susceptible-Infectious-Recovered) є класичною епідеміологічною моделлю, яка описує розповсюдження інфекційних захворювань у популяції. Модель передбачає три групи населення: *S* (Susceptible) – сприйнятливі до інфекції; *I* (Infectious) – інфіковані та здатні передавати інфекцію; *R* (Recovered) – ті, хто одужав та отримав імунітет. Клітинні автомати є дискретною моделлю, що використовує решітку клітин для моделювання просторово-часових процесів. Кожна клітина може перебувати у різних станах (наприклад, здорова, інфікована, імунна) та змінювати свій стан залежно від стану сусідніх клітин. Random Forest є методом машинного навчання, що використовує ансамбль дерев рішень для класифікації або регресії. Цей метод може бути застосований для прогнозування поширення туберкульозу на основі великої кількості вхідних параметрів. Використання цих методів дозволить провести глибокий аналіз та отримати комплексні результати щодо поширення туберкульозу на регіональному рівні в Україні. Це, в свою чергу, сприятиме розробці ефективних стратегій боротьби з хворобою та покращенню здоров'я населення.

Результати. Були детально описані та проаналізовані результати застосування методів Random Forest і SIR. Для Random Forest були оцінені метрики MSE та R^2, що показали високу точність передбачень. У випадку моделювання алгоритмом SIR, за допомогою візуальної оцінки результатів, було виявлено недостатню точність, що обумовлено недоліками моделі. Порівнюючи обрані методи з іншими дослідженнями, було зроблено висновок про необхідність розгляду більш складних алгоритмів для отримання більш точних результатів.

Висновки. На основі результатів дослідження можна зробити висновок про достатню ефективність методу Random Forest для та прогнозування уразливих соціальних груп населення та слабку ефективність алгоритму SIR для моделювання поширення туберкульозу. Для подальшого розвитку дослідження рекомендується розгляд більш складних алгоритмів та врахування додаткових факторів, що впливають на поширення захворювання. Крім того, для кращого розуміння подальших дій для поротьби з хворобою, доцільно буде провести симуляцію поширення туберкульозу серед населення України.

КЛЮЧОВІ СЛОВА: метод, метрика, туберкульоз, Random Forest, Susceptible-Infectious-Recovered, моделювання, алгоритм.

ЛІТЕРАТУРА

1. Duko B. The prevalence of depression among patients with tuberculosis: a systematic review and meta-analysis / B. Duko, A. Bedaso, G. Ayano // Ann Gen Psychiatry. – 2020. – Vol. 19(1). – P. 1. – Mode of access: https://doi.org/10.1371/journal.pone.0227472.
2. Association between tuberculosis and depression on negative outcomes of tuberculosis treatment: a systematic review and meta-analysis / [P. Ruiz-Grosso, R. Cachay, A. De La Flor et al.] // PLoS ONE. – 2020. – Vol. 15(1). – P. 1. – Mode of access: https://doi.org/10.1371/journal.pone.0227472.
3. The global prevalence of latent tuberculosis: a systematic review and meta-analysis / [A. Cohen, V.D. Mathiasen, T. Schön et al.]// Eur Respir J. – 2019. – Vol.54(3). – P. 1. – Mode of access: https://doi.org/10.1183/13993003.00655-2019.
4. Kiazyk S. Tuberculosis (TB): latent tuberculosis infection: an overview / S. Kiazyk, T. Ball // Canada Commun Dis Rep. – 2017. – Vol. 43(3–4). – P. 62. – Mode of access: https://doi.org/10.14745/ccdr.v43i34a01.
5. World Health Organization. Annual Report of Tuberculosis. Annual Global TB Report of WHO [Electronic resource]. – Access mode: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 (access date: 11.05.2024). – Title from the screen.
6. Whole genome sequencing-based analysis of tuberculosis (TB) in migrants: Rapid tools for crossborder surveillance and to distinguish between recent transmission in the host country and new importations / [E. Abascal, L. Pérez-Lago, M. Martínez-Lirola et al.] // Eurosurveillance. – 2018. – Vol. 24(4). – P. 1800005. – Mode of access: https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800005.
7. Molecular surveillance of multi- and extensively drug-resistant tuberculosis transmission in the European Union from 2003 to 2011 / [J. L. De Beer, C. Ködmön, M. J. van der Werf et al.] // Eurosurveillance. – 2014. – Vol. 19(11). – P. 20742. – Mode of access: https://doi.org/10.2807/1560-7917.ES2014.19.11.20742.
8. Whole genome sequencing of multidrug-resistant Mycobacterium tuberculosis isolates collected in the Czech Republic, 2005–2020 / [M. Dohál, V. Dvořáková, M. Šperková et al.]// Sci Rep. – 2022. – Vol. 12(1). – P. 1–10. – Mode of access: https://doi.org/10.1038/s41598-022-11287-5.
9. Anti-tuberculosis drug resistance in Slovakia, 2018–2019: the first whole-genome epidemiological study / [M. Dohál, V. Dvořáková, M. Šperková et al.] // J Clin Tuberc Other Mycobact Dis. – 2022. – Vol. 26. – P. 100292. – Mode of access: https://doi.org/10.1016/j.jctube.2021.100292.
10. Alarming levels of multidrug-resistant tuberculosis in Ukraine: results from the first national survey / [E. Pavlenko, A. Barbova, A. Hovhannesyan et al.] // Int J Tuberc

DOI 10.15588/1607-3274-2024-4-4

Lung Dis. – 2018. – Vol. 22(2). – P. 197–205. – Mode of access: https://doi.org/ 10.5588/ijtld.17.0254.
11. Vyklyuk Ya. GeoCity – a New Dynamic-Spatial Model of Urban Ecosystem / Ya. Vyklyuk, D. Nevinskyi, N. Boyko // J. Geogr. Inst. Cvijic. – 2023. – Vol. 73(2). – P. 187–203. – Mode of access: https://doi.org/10.2298/IJGI2302187V.
12. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study / [M. Zignol, A. M. Cabibbe, A.S. Dean et al.] // Lancet Infect Dis. – 2018. – Vol. 18(6). – P. 675–683. – Mode of access: https://doi.org/10.1016/S1473-3099(18)30073-2.
13. Public health concerns about Tuberculosis caused by Russia/Ukraine conflict / [L. K. Paul, G. Nchasi, D. B. Bulimbe et al.] // Health Sci Rep. – 2023. – Vol.6(4). – P. 1. – Mode of access: https://doi.org/10.1002/hsr2.1218.
14. Tuberculosis in migrants moving from high-incidence to low-incidence countries: a population-based cohort study of 519,955 migrants screened before entry to England, Wales, and Northern Ireland. / [R. W. Aldridge, D. Zenner, P. J. White et al.] // Lancet. – 2016 – Vol. 388(10059). – P. 2510. – Mode of access: https://doi.org/10.1016/S0140-6736(16)31008-X.
15. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage / [M. Merker, C. Blin, S. Mona et al.] // Nat Genet. – 2015. – Vol. 47(3). – P. 242–249. – Mode of access: https://doi.org/10.1038/ng.3195.
16. Next-generation sequencing for characterizing drug resistance-conferring mycobacterium tuberculosis genes from clinical isolates in the Ukraine / [L. T. Daum, O. S. Konstantynovska, O. S. Solodiankin et al.] // J Clin Microbiol. – 2018. – Vol. 56(6). – P. 1. – Mode of access: https://doi.org/10.1128/JCM.00009-18.
17. Mycobacterium tuberculosis Beijing lineage and risk for tuberculosis in child household contacts. / [C. C. Huang, A. L. Chu, M. C. Becerra et al.] // Peru. Emerg Infect Dis. – 2020. – Vol. 26(3). – P. 568. – Mode of access: https://doi.org/10.3201/eid2603.191314.
18. Applying geospatial multi-agent system to model various aspects of tuberculosis transmission / [Y. Vyklyuk, I. Semianiv, D. Nevinskyi et al.] // New Microbes and New Infections. – 2024. – Vol. 59. – P. 101417. – Mode of access: https://doi.org/10.1016/j.nmni.2024.101417.
19. Standard genotyping overestimates transmission of mycobacterium tuberculosis among immigrants in a low-incidence country / [D. Stucki, M. Ballif, M. Egger et al.] // J Clin Microbiol. – 2016. – Vol.54(7). – P. 1862–70. – Mode of access: https://doi.org/10.1128/JCM.00126-16.
20. Jackson S. Effects of migration on tuberculosis epidemiological indicators in low and medium tuberculosis incidence countries: a systematic review / S. Jackson, Z. Kabir, C. Comiskey // J Clin Tuberc Other Mycobact Dis. – 2021. – Vol. 23. – P. 2405–5794. – Mode of access: https://doi.org/10.1016/j.jctube.2021.100225.

DOI 10.15588/1607-3274-2024-4-4