**Kseniia BAZILEVYCH[1], Olena KYRYLENKO[1], Yurii PARFENYUK[2],
Serhii KRIVTSOV[1], Ievgen MENIAILOV[2], Victoriya KUZNIETCOVA[2],
Dmytro CHUMACHENKO[1]**

[1] *National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine*
[2] *V. N. Karazin Kharkiv National University, Kharkiv, Ukraine*

COMPARATIVE ANALYSIS OF THE MACHINE LEARNING MODELS DETERMINING COVID-19 PATIENT RISK LEVELS

*The COVID-19 pandemic has posed unprecedented challenges to global healthcare systems, emphasizing the need for predictive tools for resource allocation and patient care. This study delves into the potential of machine learning models to predict the risk levels of COVID-19 patients using a comprehensive dataset. This study **aimed** to evaluate and compare the efficacy of three distinct machine learning methodologies – Bayesian Criterion, Logistic Regression, and Gradient Boosting – in predicting the risk associated with COVID-19 patients based on their symptoms, status, and medical history. This research is **targeted** at the process of patient state determination. The research **subjects** are machine learning methods for patient state determination. To achieve the aim of the research, the following **tasks** have been formulated: methods and models of the COVID-19 patients state determination should be analyzed; classification model of the patient state determination based on Bayes criterion should be developed; classification model of the patient state determination based on logistic regression should be developed; classification model of the patient state determination based on gradient boosting should be developed; the information system should be developed; the experimental study based on machine learning methods should be provided; and the results of the experimental study should be analyzed. **Methods**: using a dataset provided by the Mexican government, which encompasses over a million unique patients with 21 distinct features, we developed an information system in C# programming language. This system allows users to select their preferred method for risk calculation, offering a real-time decision-making tool for healthcare professionals. **Results**: All models demonstrated commendable accuracy levels. However, subtle differences in their performance metrics, such as sensitivity, precision, and the F1-score, were observed. The Gradient Boosting method slightly outperformed the other models in terms of overall accuracy. **Conclusions**: While each model showcased its merits, the choice of method should be based on the specific needs and constraints of the healthcare system. The Gradient Boosting method emerged as marginally superior in this study. This research underscores the potential of machine learning in enhancing pandemic response strategies, offering both scientific insights and practical tools for healthcare professionals.*

Keywords: patient state determination; classification; machine learning; COVID-19; Bayes criterion; logistic regression; gradient boosting.

Introduction

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, emerged in late 2019 in Wuhan, China, and rapidly spread globally. Characterized by its high transmissibility and clinical manifestations ranging from asymptomatic cases to severe respiratory distress, the disease has posed unprecedented challenges to global public health, economies, and daily life [1]. The pandemic's dynamic nature, coupled with varying symptomatology and outcomes, has necessitated a multifaceted approach to its understanding and management [2].

As the world grapples with the ongoing ramifications of the pandemic, the actuality and immediacy of addressing COVID-19 remain paramount. Despite the global rollout of vaccines and the development of thera-peutic interventions, new virus variants and fluctuating infection rates underscore the persistent threat and need for innovative solutions [3].

Machine learning, a subset of artificial intelligence, has emerged as a potent tool in the fight against COVID-19. Its ability to process vast datasets [4], identify patterns [5], and make predictions has been harnessed for surveillance [6], forecasting [7], and patient management [8]. Integrating machine learning into healthcare systems has enabled real-time monitoring of the pandemic's progression and the identification of potential hotspots [9].

The task of automating diagnostics for COVID-19 using information technologies has gained significant attention [10]. Rapid, accurate diagnosis is crucial for managing the spread of the disease, and traditional diagnostic methods, such as PCR testing, have limitations

in scalability and turnaround time [11]. When trained on clinical, radiological, and laboratory data, machine learning models can expedite the diagnostic process [12].

Classification using machine learning for COVID-19 diagnostics have shown promise in differentiating between COVID-19 and other respiratory illnesses based on imaging and clinical data [13]. These models, when fine-tuned, can achieve high levels of accuracy, aiding clinicians in making informed decisions.

Furthermore, determining the condition and potential trajectory of a COVID-19 patient is of utmost importance [14]. Early identification of patients who may progress to severe stages allows for timely interventions and optimized resource allocation. Machine learning models, trained on diverse patient data, can predict outcomes and assist healthcare professionals in tailoring patient-specific care plans.

Thus, this paper aims to develop an information system based on machine learning methods for determining the state of COVID-19 patients. This research is targeted at the process of patient state determination. The research subjects are machine learning methods for patient state determination.

To achieve the aim of the research, the following tasks have been formulated:

1. Models of the COVID-19 patients' state determination should be analyzed.

2. Classification model of patient state determination based on Bayes criterion should be developed.

3. Classification model of patient state determination based on logistic regression should be developed.

4. Classification model of patient state determination based on gradient boosting should be developed.

5. An information system should be developed.

6. An experimental study based on machine learning methods should be provided.

7. The results of the experimental study should be analyzed.

The respective contributions of this research are two-fold. First, developing models based on the Bayes criterion, logistic regression, and gradient boosting methods will allow estimation of the efficiency of such methods applied to the determination of the COVID-19 patients' state. Second, the development of an information system can optimize the surveillance of healthcare working on COVID-19 patients.

In this paper, section 1, namely the current research analysis, provides the current state of COVID-19 patients' state determination methods. Section 2, Materials and Methods, provides a mathematical formulation of the task, an overview of the Bayes criterion, logistic regression, and gradient boosting methods. Section 3 provides the results of the experimental study. The dis-

cussion section discusses the obtained results. Conclusions describe the outcomes of the investigation.

Research is part of a complex intelligent information system for epidemiological diagnostics, the concept of which is discussed in [15].

1. Current Research Analysis

The study [16] investigated the heightened vulnerability of diabetic patients to COVID-19 complications, noting that while there is no concrete evidence suggesting that people with diabetes are more susceptible to contracting the virus, they face a significantly higher risk of mortality from its complications. To address this, the study introduces a risk prediction model for patients with diabetes, using a fuzzy inference system combined with machine learning techniques. The model, designed to assess COVID-19 risk without immediate medical consultation, employs eight key symptom parameters specific to individuals with diabetes. Fifteen distinct models were developed by applying multiple advanced machine learning methods. The CatBoost classifier emerged as the most effective, achieving 76% accuracy post-hyper-parameter optimization. This was closely followed by logistic regression and XGBoost, which recorded accuracies of 75.1% and 74.7%. This study employed stratified k-fold cross-validation for validation.

The paper [17] addresses the challenges faced by medical professionals in accurately predicting the prognosis of patients diagnosed with COVID-19. Recognizing the limitations of existing scoring systems in identifying patient deterioration, this study explores the potential of machine learning (ML) as an alternative. The primary objective of this study was to develop a machine learning-based method that could prospectively validate the requirement for ventilation in COVID-19 patients within a day of their initial medical encounter. This was assessed through a multicenter clinical trial involving 197 patients across five U.S. health systems from March 24 to May 4, 2020. The results indicated that the machine learning algorithm outperformed the Modified Early Warning Score (MEWS), a comparator early warning system, in predicting ventilation needs. Specifically, the algorithm showcased a diagnostic odds ratio (DOR) of 12.58 and a sensitivity of 0.90, surpassing MEWS's sensitivity of 0.78. The study concludes that the machine learning algorithm offers a promising tool for accurately predicting the need for mechanical ventilation in COVID-19 patients within 24 h, potentially aiding in better patient triage and resource allocation.

The study [18] examined the impact of ventilator-associated pneumonia (VAP) on the outcomes of mechanically ventilated patients, particularly those with severe COVID-19. Although guidelines emphasize VAP

prevention and treatment, its role in patient outcomes remains ambiguous. This study assessed the link between unsuccessful VAP treatment and mortality in patients with severe pneumonia. Conducted as a single-center, prospective cohort study, this study involved 585 patients with severe pneumonia and respiratory failure, of whom 190 were diagnosed with COVID-19. All participants underwent at least one bronchoalveolar lavage. To understand the extended intensive care unit (ICU) stay observed in COVID-19 patients, a machine-learning method named CarpeDiem was employed. This approach categorized similar ICU patient days into clinical states using electronic health record data. The findings indicated that prolonged ICU stays for COVID-19 patients were primarily due to extended respiratory failure. Although VAP was not directly linked to overall mortality, patients with one episode of unsuccessfully treated VAP had a significantly higher mortality rate than those with successful VAP treatment (76.4% vs. 17.6%). CarpeDiem further highlighted that unresolved VAP led to transitions into clinical states with increased mortality risks. In conclusion, this study underscores the association between unsuccessful VAP treatment and elevated mortality, emphasizing the heightened VAP risk in COVID-19 patients due to prolonged respiratory failure.

The research [19] employs multivariate and machine learning techniques to predict mortality rates among hospitalized COVID-19 patients. Using data from 250 patients, a model based on the "Inspired modification of partial least square (SIMPLS)" was developed, showcasing high accuracy in differentiating survivors from non-survivors. Key mortality predictors included specific clinical features, comorbidities such as coronary artery disease and diabetes, and blood markers such as CRP and prothrombin. This study highlights the potential of machine-learning models in enhancing COVID-19 patient management by accurately predicting outcomes and identifying high-risk individuals.

The study [20] employed a machine learning algorithm to pinpoint a specific subset of COVID-19 patients who might benefit from hydroxychloroquine treatment in terms of survival. Conducted across six U.S. hospitals, this pragmatic trial involved 290 COVID-19 patients admitted between 10 March and June 4, 2020 without randomized treatment. The primary outcome was mortality. The algorithm identified a subpopulation in which hydroxychloroquine significantly improved survival rates (p=0.011), with an adjusted survival rate of 82.6% in the treated group compared with 51.2% in the untreated group. However, this association was not observed in the broader patient population. Precision medicine techniques can help identify specific COVID-19 patient groups that may benefit from hydroxychloroquine in clinical trials.

In the contemporary medical research landscape, machine learning has emerged as an indispensable tool for patient state determination, particularly in the COVID-19 pandemic. A synthesis of recent scholarly investigations reveals a consistent trend: machine learning models exhibit superior precision in predicting intricate patient outcomes, often surpassing conventional methodologies. For instance, one seminal study elucidated the efficacy of a machine learning algorithm in discerning a distinct cohort of patients with COVID-19 who demonstrated enhanced survival outcomes with hydroxychloroquine administration. Such granularity in patient stratification is paramount, especially in scenarios characterized by resource constraints or where bespoke interventions can significantly alter patient trajectories. Furthermore, the adeptness of machine learning in deciphering multifaceted clinical phenotypes provides an unparalleled depth of insight and is instrumental for nuanced patient management. Given these empirical findings, our endeavor to harness machine learning for patient state determination is justified. It emerges as a quintessential progression, heralding a paradigm shift toward more nuanced, individualized, and effective patient care.

2. Materials and Methods

2.1. Mathematical Formulation of the Classification Task

Let us consider a set M of objects ω on this set, a partition exists into a finite number of subsets (classes):

$$\Omega_i, i = 1, \ldots, m, M = \bigcup_{i=1}^{m} \Omega_{i_i}. \qquad (1)$$

The partition of M is not fully defined. Only some information I_0 about class Ω_i is given. Objects ω are defined by the values of certain features x_j, j=1, …, p (this set is always the same for all objects considered when solving a specific problem). The set of feature values x_j defines the description $I(\varpi)$ of the object ϖ. Each feature can take values from different sets of permissible feature values, for example, from the following: (0, 1) – the feature is absent or present, respectively; (0, 1, Δ), Δ – information about the feature is missing; (0, 1, …, d − 1) – the degree of expressiveness of the feature has different gradations, d>2; a_1, …, a_d – the feature has a finite number of values, d>2; functions of a specific class define the values of the feature x_j; the values of the feature x_j are the distribution functions of a particular random variable. An object's $I(\omega) = (x_1(\varpi), \ldots, x_n(\varpi))$ description is standard if $x_j(\varpi)$ takes values from the set of permissible values.

The classification task with standard information is to compute the values of predicates $P_i(\varpi) = "\varpi \in \Omega_i"$,

i = 1, …, m for a given object ϖ and a set of classes Ω_i, I = 1, …, m based on the training information I_0 (Ω_i, …, Ω_m) about the classes and the description I (ϖ). Information about the inclusion of the object ϖ in the class Ω_i is encoded by the symbols "1"(ϖ ∈ Ω_i), "0"(ϖ ∉ Ω_i), Δ – it is unknown whether ω belongs to the class Ω_i or not [21].

2.2. Determining the Patient's Condition and Making a Diagnosis as a Classification Task

Diagnosis is a classification task [22]. A medical diagnosis can be both a designation of a pathological state of the human body and an identification of the causative disease that induced this state. In the former case, it pertains to classify human body states that need to be diagnosed on the basis of particular descriptions. At the same time, the latter concerns the classification of causes and the diagnosis of reasons that led to changes in the body's state. The classification of human diseases can structurally take the form of a tree, with its terminal nodes representing diagnoses.

The diagnostic process is visualized as traversing this tree, contingent on answers to questions posed at each node. In reality, experts cannot obtain direct answers during diagnosis and are confined to indirect queries. Therefore, the disease classification structure can only serve diagnostic purposes to a limited extent, emphasizing the importance of developing diagnostic features and constructing a rule [23]. These features and the diagnostic rule should be derived from educational statistical data.

Classification algorithms used in medical diagnostics are grounded in three hypotheses. The first posits that, given a comprehensive description, similar descriptions in the feature space should correspond to proximate diagnoses. The second hypothesis postulates a functional type of class separability, such as linear separability, quadratic separability, etc. The third hypothesis assumes the existence of a diagnostic combination of features that appear in one class much more frequently than in another.

In this study, three methods were chosen to address the research task: the Bayes criterion, logistic regression, and gradient boosting.

2.3. Bayes Criterion

he Bayes criterion is a statistical classification method used for decision-making regarding classifying an object into a specific class based on its features or characteristics [24]. This method applies Bayes' theorem from probability theory [25].

The core idea of the Bayes criterion is to determine the probabilities of an object belonging to each class based on the input data. These probabilities are calculated using prior class probabilities and the likelihood functions of the features for each class. Based on the application of Bayes' theorem, the concept of Bayesian data analysis anticipates its recurrent use upon the emergence of new measurements, facts, and expert evaluations. The fundamental rule for the calculation is as follows:

$$p(A|B) * p(B) = p(A \cap B). \qquad (2)$$

The probability that both events A and B will occur is determined by the probability of A given B and the probability of B. Introducing a condition regarding another event leads to the expression:

$$p(A|B \cap C) * p(B|C) = p(A \cap B|C). \qquad (3)$$

Since $p(A \cap B) = p(B \cap A)$, then $p(A \cap B|C) = p(B \cap A|C)$. This leads to Bayes' theorem. Bayes' rule:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \qquad (4)$$

Bayes' rule is a method for updating the probability of event A given that there is information about event B, where p(A) is the prior probability of event A, p(A|B) is the posterior probability of event A given that event B occurred, and p(B|A) is the likelihood of event A given event B. The formula for the Bayes criterion is:

$$p(A|x) = \frac{p(x|A)p(A)}{p(x)}. \qquad (5)$$

where P(A|x) is the probability of an object belonging to class A given the observed features x, P(x|A) is the likelihood function of feature x for class A, P(A) is the prior probability of class A, and P(x) is the probability of feature x.

The conditional advantages of the method identified compared with other research methods include:

1. The Bayes criterion considers the complexity of the model.

2. It provides a fundamental way to select the best model from the alternatives.

3. The use of prior knowledge can be beneficial when working with small datasets.

Drawbacks include:

1. Intensive computations due to the need for sampling or numerical integration.

2. Assumptions about prior distributions can introduce bias if the prior distribution is incorrectly defined.

3. The performance of Bayesian methods largely depends on the choice of prior distributions.

2.4. Logistic Regression

Logistic regression is a statistical method primarily used for binary classification problems [26]. In the context of patient classification, it is a powerful tool for predicting the probability of a patient belonging to a particular category based on one or more predictor variables. Unlike linear regression, which predicts a continuous outcome, logistic regression focuses on predicting the probability that a given instance belongs to a specific category.

The essence of logistic regression lies in its use of the logistic function, which is often called the sigmoid function. This function models the relationship between a binary dependent variable and one or more independent variables. Mathematically, the logistic function can be represented as:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}, \qquad (6)$$

where $P(Y=1)$ is the probability of the dependent event occurring, β_0 is the intercept, and $\beta_1, \ldots, \beta_k$ are the coefficients of the predictor variables $x_1, \ldots, x_k$.

The coefficients in the logistic regression model are typically estimated from the data using the maximum likelihood method. Once trained, the model can estimate the probability of a new patient falling into a particular category on the basis of their predictor variables. A threshold, commonly set at 0.5 for binary classification, is then applied to these probabilities to classify patients into the respective categories.

Advantages:

1. Logistic regression is relatively straightforward to interpret. The coefficients of the predictor variables indicate the change in log odds of the outcome, making it easy to explain the influence of the predictor variables.

2. Logistic regression provides probabilities that the given input point belongs to a particular category, which is crucial for understanding the confidence of the prediction.

3. Being a linear model, logistic regression tends to have low variance and is less prone to overfitting, especially when regularized.

4. Logistic regression is computationally less intensive than more complex models, making it a good choice for problems with many features or data points.

Disadvantages:

1. While logistic regression does not assume linearity between variables and outcome, it does assume linearity between independent variables and log odds.

2. Logistic regression might not perform well when non-linear boundaries separate the classes. More complex algorithms such as decision trees or neural networks might better suit such cases.

3. Logistic regression requires that the independent variables are not highly correlated, a condition known as multicollinearity. If present, it can result in unreliable and unstable estimates of regression coefficients.

4. Logistic regression is sensitive to outliers in the dataset, which can adversely affect the performance of the model.

2.5. Gradient Boosting

Boosting is a class of machine learning methods founded on the idea that a combination of simple classifiers (obtained by a weak learner) can perform better than any individual simple classifier [27]. Gradient boosting, or Gradient Boosting Classifier, is a machine learning method for regression and classification tasks that constructs a predictive model in the form of an ensemble of weak predictive models, typically decision trees [28]. This method iteratively adds models to the ensemble. It is used to minimize bias and variance in supervised learning.

The gradient boosting algorithm is described as follows:

1. The model is constructed on a subset of data.

2. Predictions are calculated on this subset of data.

3. Errors are computed by comparing the predictions and actual values.

4. A new model is created using the errors as the target variable. The aim is to find the best split to minimize the error.

5. The predictions made by this new model are combined with the predictions of the previous model using formula 1.8 (see below).

6. New errors are calculated using this predicted value and the actual value.

7. Steps 2-6 are repeated until the maximum number of iterations is reached (or the error function remains unchanged).

Elements of gradient boosting include:

1. A loss function to be optimized: This function depends on the problem being solved. It must be differentiable, but many standard loss functions are supported, and custom functions can be defined.

2. A "weak learner" (WL) is a learning algorithm that produces classifiers with errors that are strictly less than random guessing. In contrast, a "strong learner" (SL) learns (by training on sufficient training data) and gives classifiers with persistently low error probability. Decision trees are used as the "weak learner" in gradient boosting.

3. An additive model to add "weak learners" to minimize the loss function. Trees are added one at a time, and existing trees in the model are not changed. The gradient descent procedure is used to minimize loss when adding trees.

The gradient boosting formula is:

$$F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^{j_m} \gamma_{jm} I(x \in R_{jm}), \quad (7)$$

where $F_m(x)$ is the m-th extended tree model, η is the learning rate, $I(x \in R_{jm})$ is the indicator function equal to 1 if $x \in R_{jm}$, and 0 otherwise, and γ_{jm} is the loss function minimizer.

The conditional advantages of the method, identified compared with other research methods, include:

1. High prediction accuracy, often surpassing other algorithms in many domains.

2. Capability to work with heterogeneous types of data, including categorical and numerical features.

3. Ability to capture complex interactions and non-linear relationships.

Disadvantages include:

1. It is more computationally intensive and slower to train than simpler models.

2. Interpretation can be challenging because of the complexity of the ensemble model.

3. Results

3.1. Data

The input data are presented in Table 1.

Table 1

The input data

Name	Type	Range
Treatment in medical institutions	Boolean	0, 1
Medical insurance	Integer	[1–12]
Gender	Boolean	0, 1
Patient type	Boolean	0, 1
Pneumonia	Boolean	0, 1
Age	Integer	[30–121]
Diabetes	Boolean	0, 1
Chronic lung disease	Boolean	0, 1
Asthma	Boolean	0, 1
Weakened immunity	Boolean	0, 1
Elevated blood pressure	Boolean	0, 1
Other disease	Boolean	0, 1
Cardiovascular disease	Boolean	0, 1
Obesity	Boolean	0, 1
Chronic kidney disease	Boolean	0, 1
Smoking	Boolean	0, 1
Illness	Boolean	0, 1

The output data are presented as decimal numbers.

The experimental study was conducted using data from an open dataset [29]. The "COVID-19 Dataset" on Kaggle provides anonymized patient data related to the COVID-19 pandemic, sourced from the Mexican government. It encompasses information on more than a million unique patients across 21 distinct attributes. The dataset's primary objective is to facilitate the development of machine learning models that can predict a patient's risk level based on symptoms, status, and medical history. Key features include demographic data (e.g., age, sex), medical history (e.g., diabetes, hypertension), and COVID-19-specific data (e.g., classification of test findings, patient type). This dataset aids healthcare providers in efficiently allocating medical resources by predicting patient needs.

3.2. Information System

An information system was developed using the C# programming language to automate the classification process. The user must select the method they wish to use to calculate the probability of illness by clicking on the method's name. Figure 1 provides an overview of the method selection.

Fig. 1. Selection of the method for calculation probability of illness

Upon pressing the "Calculate" button, the probability of contracting COVID-19 is computed. Figure 2 displays the results of the calculation.
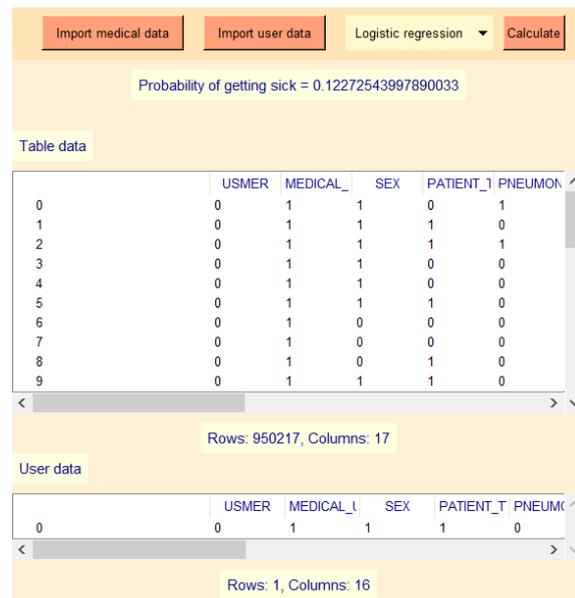

Fig. 2. Calculation of the examined patient's condition

3.3. Experimental Study

Let us consider the results of the information system using the data of a specific patient as an example (Table 3).

Initially, we identify the correlation matrix and determine each table column's maximum and minimum values. This is a necessary condition to determine the applicability of the criterion. Table 4 provides an overview of the correlation matrix.

Table 3

Data of the patient

Name	Input data
Treatment in medical institutions	0
Medical insurance	1
Gender	1
Patient type	1
Pneumonia	0
Age	56
Diabetes	0
Chronic lung disease	0
Asthma	0
Weakened immunity	0
Elevated blood pressure	0
Other disease	0
Cardiovascular disease	1
Obesity	1
Chronic kidney disease	0
Smoking	0

Table 4

Classification matrix

Variable	Value	Value
USMER	0.1173	-0.15071
MEDICAL_UNIT	0.16479	-0.11951
SEX	0.060682	-0.10055
PATIENT_TYPE	0.16479	-0.59429
PNEUMONIA	0.17393	-0.59429
AGE	0.36671	-0.20162
DIABETES	0.35595	-0.19311
COPD	0.14943	-0.096475
ASTHMA	0.04609	-0.021778
INMSUPR	0.12004	-0.086594
HYPERTENSION	0.36671	-0.16616
OTHER_DISEASE	0.12004	-0.11481
CARDIOVASCULAR	0.1485	-0.08496
OBESITY	0.16476	-0.045375
RENAL_CHRONIC	0.16099	-0.12353
TOBACCO	0.78432	-0.10055
CLASSIFICATION	0.11492	-0.11557

Let us examine the Bayes criterion implemented in the information system to determine the probability of illness.

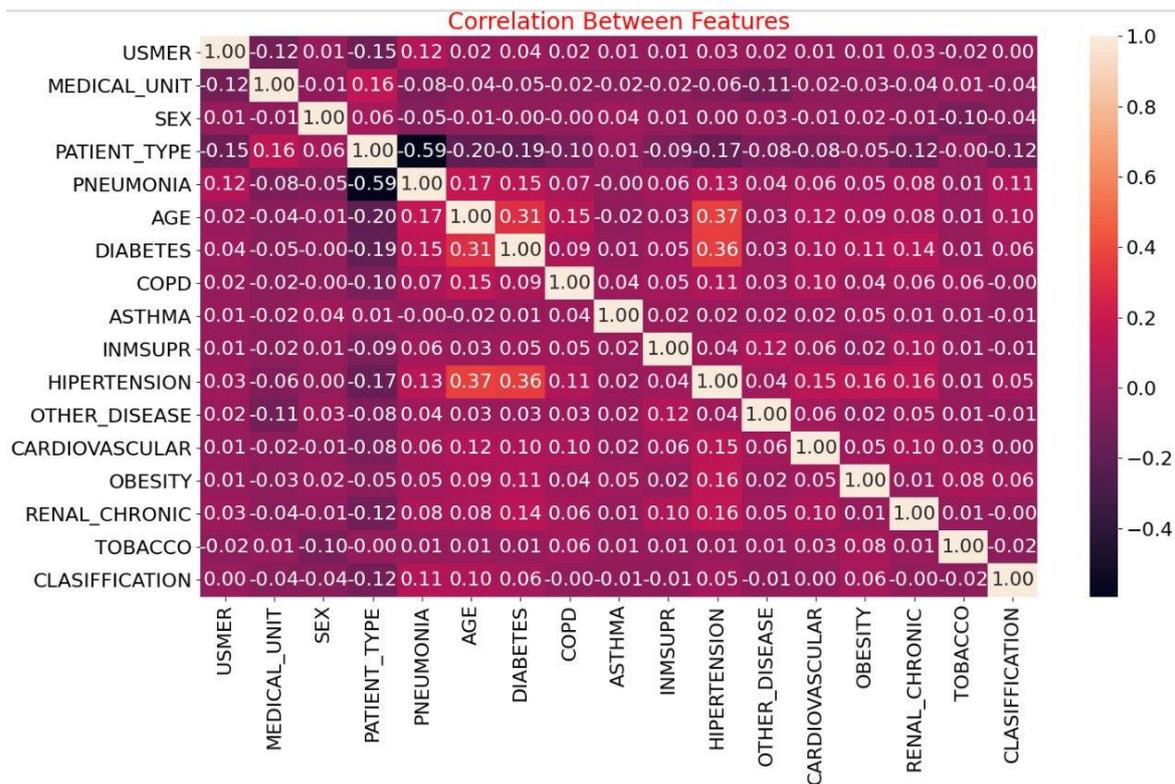Figure 3 presents a visualization of the correlation matrix in the form of a heatmap:

Fig. 3. Correlation matrix

In our research, the correlation matrix is a pivotal tool for understanding the linear relationships between the input variables. By analyzing these relationships, we can identify potential multicollinearity that could adversely affect the performance and interpretability of our machine learning models. The correlation matrix was compiled using Pearson's correlation coefficient, quantifying the linear relationship between the two variables. Each entry in the matrix represents the correlation coefficient between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value close to 0 indicates a weak correlation. By examining this matrix, we can make informed decisions about feature selection and potential modifications to our modeling approach, thereby ensuring the robustness and reliability of our results.

For the given sample, the interval is [-0.59; 0.366], indicating a low correlation dependence between the input variables. Therefore, the criterion can be applied to this dataset.

To calculate using the criterion itself, maximizing the product of probabilities $P(X|C_k)P(C_k)$ is necessary for k = 2, as there are only two classes in this task. The a priori probability of the "healthy" class is calculated using the following formula:

$$P(C_1) = \frac{614769}{950217} = 0.65. \qquad (8)$$

For "infected" class:

$$P(C_2) = \frac{335448}{950217} = 0.35. \qquad (9)$$

Let's calculate the conditional possibilities for the data (Table 5).

The results of the conditional probabilities are:

$$P(X|C_1) = 3.018E-09, \qquad (10)$$
$$P(X|C_2) = 7.34E-09. \qquad (11)$$

Then, the probabilities $P(X|C_k)P(C_k)$ will respectively be:

$$P(X|C_1)P(C_1) = 1.95E-09, \qquad (12)$$
$$P(X|C_2)P(C_2) = 2.58E-09. \qquad (13)$$

Next, we normalize the result:

$$\frac{P(X|C_1)P(C_1)}{P(X|C_1)P(C_1)+P(X|C_2)P(C_2)} = 0.7115, \qquad (14)$$

$$\frac{P(X|C_2)P(C_2)}{P(X|C_1)P(C_1)+P(X|C_2)P(C_2)} = 0.2885. \qquad (15)$$

Consequently, a patient with the described characteristics has a probability of 0.2885 of being diagnosed as sick (falling into the "Infected" class), and a 0.7115 probability of being diagnosed as healthy (falling into the "Healthy" class).

Table 5

Conditional probabilities

Description	Calculation
$P(Usmer=0|C_1)$	405350/614769=0.66
$P(Usmer=0|C_2)$	219989/335448=0.65
$P(Medical\_unit=1|C_1)$	109/614769=0.0001
$P(Medical\_unit=1|C_2)$	37/335448=0.0001
$P(Sex=1|C_1)$	323914/614769=0.526
$P(Sex=1|C_2)$	162541/335448=0.48
$P(Patient\_type=1|C_1)$	552350/614769=0.89
$P(Patient\_type=1|C_2)$	274084 /335448=0.81
$P(Pneumonia=0|C_1)$	574807/614769=0.93
$P(Pneumonia=0|C_2)$	290652/335448=0.86
$P(Age=56|C_1)$	561634/614769=0.91
$P(Age=56|C_2)$	293881 /335448=0.87
$P(Diabetes=0|C_1)$	607852/614769=0.91
$P(Diabetes=0|C_2)$	331888/335448=0.98
$P(Copd=0|C_1)$	594995/614769=0.98
$P(Copd=0|C_2)$	326172/335448=0.97
$P(Asthma=0|C_1)$	607016/614769=0.96
$P(Asthma=0|C_2)$	332112/335448=0.99
$P(Inmsupr=0|C_1)$	539954/614769=0.98
$P(Inmsupr=0|C_2)$	282222/335448=0.84
$P(Hypertension=0|C_1)$	599373/614769=0.97
$P(Hypertension=0|C_2)$	328155/335448=0.841
$P(Other\_disease=0|C_1)$	10206/614769 = 0.01
$P(Other\_disease=0|C_2)$	5594/335448=0.01
$P(Cardiovascular=1|C_1)$	131/614769=0.016
$P(Cardiovascular=1|C_2)$	59161/335448=0.17
$P(Obesity=1|C_1)$	80932/614769=0.13
$P(Obesity=1|C_2)$	331086/335448=0.98
$P(Renal\_chronic=0|C_1)$	606348/614769=0.98
$P(Renal\_chronic=0|C_2)$	311232/335448=0.92
$P(Tobacco=0|C_1)$	562730/614769=0.91
$P(Tobacco=0|C_2)$	330321/335448=0.98

In Figure 4, we observe the probability computation, which aligns with the calculation above.

Figure 5 shows the probability calculations using logistic regression.

Figure 6 shows the probability calculations using gradient boosting.

Meanwhile, the software user can independently select the calculation method depending on the research objective and other conditions.

The results demonstrate that the classification quality is high (Table 6). The following metrics were chosen to evaluate the classification quality: accuracy, sensitivity, precision, and the F1-score. Accuracy represents the ratio of correctly classified instances to the total number of instances in the test dataset. A higher accuracy indicates better classification quality. Sensitivity determines the percentage of correctly classified positive instances. High sensitivity suggests the model's proficient ability to recognize both instances. Precision reflects the percentage of correctly classified positive instances out of all positive classifications. The F1-score is the harmonic mean between precision and sensitivity, providing a balanced assessment of classification quality.
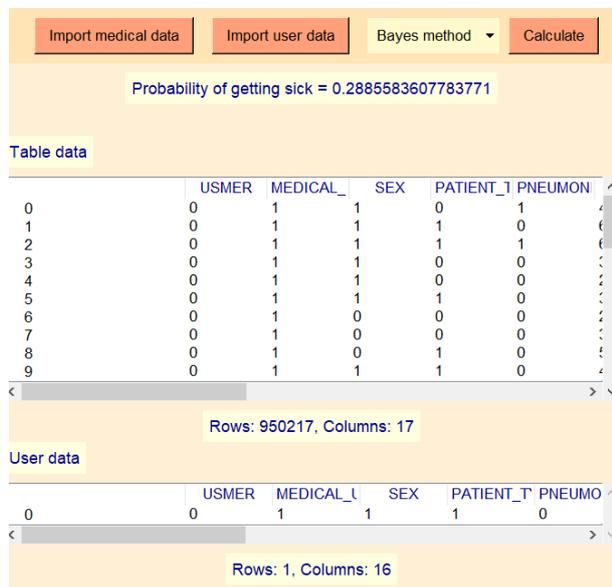

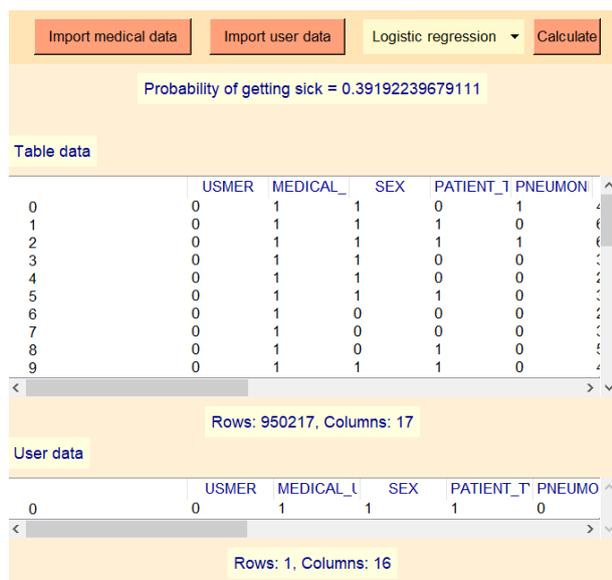Fig.4. Result of the method for the given parameters (using the Bayesian criterion)

Fig.5. Result of the method for the given parameters (using the logistic regression)
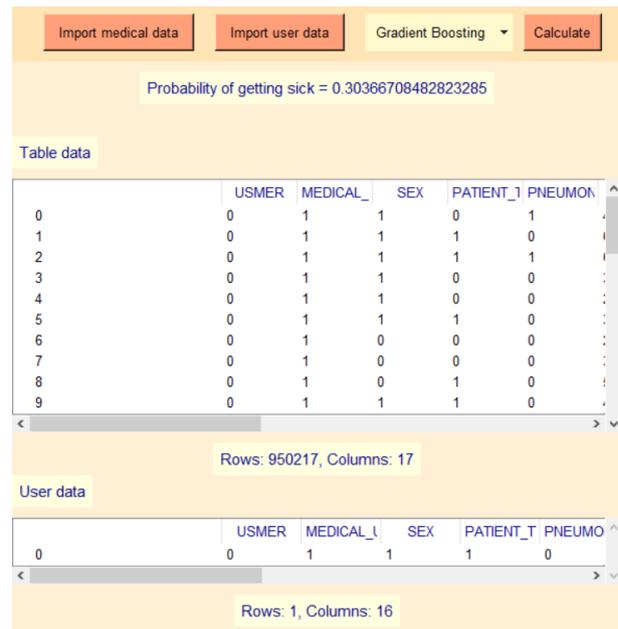

Fig.6. Result of the method for the given parameters (using the gradient boosting)

Table 6

Comparative Characteristics
of Classification Quality

Metric	Bayes Criterion	Logistic regression	Gradient boosting
Accuracy	0.6377	0.659	0.665
Sensitivity	0.285	0.117	0.1452
Precision	0.473	0.57	0.594
F1-score	0.356	0.194	0.233

4. Discussion

The COVID-19 pandemic has presented unprecedented challenges to healthcare systems worldwide. One of the primary concerns has been the efficient allocation of medical resources, given the surge in patients and potential resource shortages. Predictive models that can forecast a patient's risk level based on symptoms, status, and medical history can be invaluable tools for healthcare providers, enabling them to make informed decisions about resource allocation and patient care.

In this study, we explored the potential of various machine learning models to predict a patient's risk level for COVID-19 based on a comprehensive dataset provided by the Mexican government. This dataset, comprising more than a million unique patients, offers a rich source of information, including various patient-related factors and pre-conditions.

Our analysis began with the development of a C# programming language information system to automate classification. This system allows users to select their preferred method for calculating the probability of dis-

ease, offering a user-friendly interface for method selection and result visualization.

Three primary methods were evaluated: the Bayesian Criterion, Logistic Regression, and Gradient Boosting. Each method offers unique advantages and potential limitations. The Bayesian approach, grounded in probability theory, provides a straightforward and interpretable way to calculate the likelihood of an event based on prior knowledge. Logistic Regression, a staple in binary classification, models the probability of an event occurring on the basis of one or more predictor variables. Gradient Boosting, an ensemble method, builds a predictive model by combining the outputs of multiple weak learners, typical decision trees.

Our results indicate that all three methods offer reasonably high classification quality. The Gradient Boosting method slightly outperformed the other two methods in accuracy, achieving an accuracy of 0.665. However, its sensitivity, which measures the ability to identify positive cases correctly, was relatively low at 0.1452. The Bayesian Criterion, while achieving a moderate accuracy of 0.6377, had a higher sensitivity of 0.285. This suggests that while Gradient Boosting might improve overall classification, the Bayesian approach might be more adept at identifying positive cases, which is crucial in healthcare.

The importance of other metrics such as precision and the F1-score is also worth noting. The F1-score, which provides a balanced measure between precision and sensitivity, was highest for Gradient Boosting, followed by the Bayesian Criterion and Logistic Regression. This further underscores the potential efficacy of Gradient Boosting in this context.

However, while these results are promising, it is essential to consider their potential limitations. The dataset, though extensive, represents patients from Mexico and may not be generalizable to other populations. Additionally, missing data, represented by values such as 97 and 99 in the dataset, could introduce biases or inaccuracies in the predictions.

Another limitation is that while our software demonstrates the potential of machine learning models in predicting COVID-19 patient risk levels, its current design is primarily for research and academic exploration. The proposed system does not fully adhere to the stringent regulatory standards set forth for medical devices. This regulatory gap underscores the limitations of the software in handling, storing, and transmitting electronic medical records securely and reliably. Consequently, its applicability in real-world clinical settings remains constrained until these regulatory and data management challenges are comprehensively addressed.

In conclusion, machine learning models hold significant promise in aiding the fight against COVID-19 by predicting patient risk levels. While each method has

its strengths and weaknesses, the choice of model should be based on the application's specific requirements, whether it is overall accuracy, sensitivity, or other factors. Future research could explore the integration of these models into real-world healthcare systems, assess their performance in diverse populations, and refine them based on feedback from healthcare professionals.

Conclusions

The COVID-19 pandemic underscored the importance of developing advanced predictive tools to assist healthcare systems in making informed decisions about resource allocation and patient care. In this study, we used a comprehensive dataset provided by the Mexican government to evaluate the efficacy of various machine learning models in predicting the risk levels associated with COVID-19 patients.

From a scientific perspective, the novelty of this research lies in its unique juxtaposition of three distinct machine learning methodologies: the Bayesian Criterion, Logistic Regression, and Gradient Boosting. This comparative analysis, especially with such a vast dataset, enriches the existing literature by offering more profound insights into the relative strengths and applicabilities of these methods in the context of a global pandemic.

Furthermore, the research illuminated the nuances of each method's performance metrics. While all models demonstrated a commendable level of accuracy, subtle differences in sensitivity, precision, and the F1 score provided valuable insights into their respective strengths and potential areas of improvement.

On the practical front, the development of an information system, as illustrated, offers a user-friendly interface that allows healthcare professionals to select a preferred method for risk calculation. Such a system can be instrumental in real-time decision-making, enabling timely interventions and optimizing resource utilization.

In conclusion, while each evaluated model has its merits, the Gradient Boosting method slightly outperformed the others in terms of overall accuracy. However, the choice of method should be contingent on the specific requirements and constraints of the healthcare system in question. This research, with its dual focus on scientific exploration and practical application, paves the way for more nuanced and context-specific predictive tools in the future.

Contributions of authors: conceptualization – **Kseniia Bazilevych, Olena Kyrylenko, Serhii Krivtsov**; methodology – **Kseniia Bazilevych, Serhii Krivtsov**; formulation of tasks – **Yurii Parfenyuk, Dmytro Chumachenko**; analysis – **Yurii Parfenyuk, Serhii Krivtsov, Ievgen Meniailov, Victoriya**

Kuznietcova; development of model – **Olena Kyrylenko, Serhii Krivtsov**; software – **Olena Kyrylenko, Serhii Krivtsov, Ievgen Meniailov**; verification – **Kseniia Bazilevych, Yurii Parfenyuk, Victoriya Kuznietcova**; analysis of results – **Kseniia Bazilevych, Serhii Krivtsov, Dmytro Chumachenko**; visualization – **Olena Kyrylenko, Serhii Krivtsov**; writing – original draft preparation – **Kseniia Bazilevych, Olena Kyrylenko, Serhii Krivtsov, Dmytro Chumachenko**; writing – review and editing – **Yurii Parfenyuk, Ievgen Meniailov, Victoriya Kuznietcova**.

All authors have read and agreed with the published version of this manuscript.

References

1. Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, 2020, vol. 19, iss. 19, pp. 1–14. DOI: 10.1038/s41579-020-00459-7.

2. Esakandari, H., Nabi-Afjadi, M., Fakkari-Afjadi, J., Farahmandian, N., Miresmaeili, S.-M., & Bahreini, E. A Comprehensive Review of COVID-19 Characteristics. *Biological Procedures Online*, 2020, vol. 22, iss. 1, article no. 19. DOI: 10.1186/s12575-020-00128-2.

3. Forchette, L., Sebastian, W., & Liu, T. A Comprehensive Review of COVID-19 Virology, Vaccines, Variants, and Therapeutics. *Current Medical Science*, 2021, vol 41, iss. 6, pp. 1037-1051. DOI: 10.1007/s11596-021-2395-1.

4. Tang, R., Jiang, J., Zhang, Y., & Luo, J. Open Government Data (OGD) sites and the sharing of country-specific real-time pandemic information: An investigation into COVID-19 datasets available on worldwide OGDs. *Information Processing and Management*, 2023, vol. 60, iss. 6, pp. 103489-103489. DOI: 10.1016/j.ipm.2023.103489.

5. Zakharchenko, O., Avramenko, R., Zakharchenko, A., Korobchuk, A., Fedushko, S., Syerov, Y., & Trach, O. Multifaceted Nature of Social Media Content Propagating COVID-19 Vaccine Hesitancy: Ukrainian Case. *Procedia Computer Science*, 2022, vol. 198, pp. 682–687. DOI: 10.1016/j.procs.2021.12.306.

6. Izonin, I., Tkachenko, R., Dronyuk, I., Tkachenko, P., Gregus, M., & Rashkevych, M. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. *Mathematical Biosciences and Engineering*, 2021, vol. 18, iss. 3, pp. 2599–2613. DOI: 10.3934/mbe.2021132.

7. Badi, H., Badi, I., Moutaouakil, K. E., Khamjane, A., & Bahri, A. Sentiment analysis and prediction of polarity vaccines based on Twitter data using deep NLP techniques. *Radioelectronic and Computer Systems*, 2022, vol. 104, iss. 4, pp. 19–29. DOI: 10.32620/reks.2022.4.02.

8. Strilets, V., Donets, V., Ugryumov, M., Artiuch, S., Zelenskyi, R., & Goncharova, T. Agent-oriented data clustering for medical monitoring. *Radioelectronic and computer systems*, 2022, vol. 101, iss. 1, pp. 103–114. DOI: 10.32620/reks.2022.1.08.

9. Lukas, H., Xu, C., Yu, Y., & Gao, W. Emerging Telemedicine Tools for Remote COVID-19 Diagnosis, Monitoring, and Management. *ACS Nano*, 2020, vol. 14, iss. 12, pp. 16180–16193. DOI: 10.1021/acsnano.0c08494.

10. Arefiev, V., Kovalenko, G., Frant, M., Chumachenko, T., Polyvianna, Y., Pivnenko, S., Bolotin, V., Mayboroda, O., Solodiankin, O., Tarasov, O., Bezyemenni, M., Lyon, C., Redlinger, M., Sapachova, M., Mezhenskyi, A. A., Ducluzeau, A.-L., Bortz, E., Gerilovych, A., & Drown, D. M. Complete Genome Sequence of *Salmonella enterica* subsp. *enterica* Serovar Kottbus Strain Kharkiv, Isolated from a Commercial Pork Production Facility in Ukraine. *Microbiology Resource Announcements*, 2020, vol. 9, iss. 49, e01171-20. DOI: 10.1128/mra.01171-20.

11. Bazilevych, K., Krivtsov, S., & Butkevych, M. Intelligent Evaluation of the Informative Features of Cardiac Studies Diagnostic Data using Shannon Method. *CEUR Workshop Proceedings*, 2021, vol. 3003, pp. 65-75.

12. Meniailov, I., & Padalko, H. Application of Multidimensional Scaling Model for Hepatitis C Data Dimensionality Reduction. *CEUR Workshop Proceedings*, 2022, vol. 3348, pp. 33–43.

13. Radutniy, R., Nechyporenko, A., Alekseeva, V., Titova, G., Bibik, D., & Gargin, V. V. Automated Measurement of Bone Thickness on SCT Sections and Other Images. *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, 2020, pp. 222-226, DOI: 10.1109/dsmp47368.2020.9204289.

14. Aswathy, A. L., Anand, H. S., & Chandra, S. S. V. COVID-19 severity detection using machine learning techniques from CT-images. *Evolutionary Intelligence*, 2022, vol. 1, iss. 9, pp. 1423-1431. DOI: 10.1007/s12065-022-00739-6.

15. Yakovlev, S., Bazilevych, K., Chumachenko, D., Chumachenko, T., Hulianytskyi, L., Meniailov, I., & Tkachenko, A. The Concept of Developing a Decision Support System for the Epidemic Morbidity Control. *CEUR Workshop Proceedings*, 2020, vol. 2753, pp. 265–274.

16. Aggarwal, A., Chakradar, M., Bhatia, M. S., Kumar, M., Stephan, T., Gupta, S., Alsamhi, S. H., & Al-Dois, H. COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine

Learning Approaches. *Journal of Healthcare Engineering*, 2022, vol. 2022, pp. 1–10. DOI: 10.1155/2022/4096950.

17. Burdick, H., Lam, C., Mataraso, S., Lynn-Palevsky, A., Braden, G., Dellinger, R.P., McCoy, A., Vincent, J.-L., Green-Saxena, A., Barnes, G., Hoffman, J., Calvert, J., Pellegrini, E., & Das, R. Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Computers in Biology and Medicine*, 2020, vol. 124, article no. 103949. DOI: 10.1016/j.compbiomed.2020.103949.

18. Gao, C.A., Markov, N. S., Stoeger, T., Pawlowski, A. E., Kang, M., Nannapaneni, P., Grant, R. A., Pickens, C., Walter, J. M., Kruser, J.M., Rasmussen, L. V., Schneider, D., Starren, J., Donnelly, H. K., Donayre, A., Luo, Y., Budinger, S., Wunderink, R. G., Misharin, A. V., & Singer, B. D. Machine learning links unresolving secondary pneumonia to mortality in patients with severe pneumonia, including COVID-19. *Journal of Clinical Investigation*, 2023, vol. 133, iss. 12, article no. e170682. DOI: 10.1172/jci170682.

19. Banoei, M. M., Dinparastisaleh, R., Zadeh, A. V., & Mirsaeidi, M. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. *Critical Care*, 2021, vol. 25, iss. 1, article no. 328. DOI: 10.1186/s13054-021-03749-5.

20. Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R. P., McCoy, A. J., Vincent, J.-L., Green-Saxena, A., Barnes, G. L., Hoffman, J., Calvert, J., Pellegrini, E., & Das, R. Is Machine Learning a Better Way to Identify COVID-19 Patients Who Might Benefit from Hydroxychloroquine Treatment? – The IDENTIFY Trial. *Journal of Clinical Medicine*, 2020, vol. 9, iss. 12, pp. 3834-3834. DOI: 10.3390/jcm9123834.

21. Mashtalir, V. P., Shlyakhov, V. V., & Yakovlev, S. V. Group Structures on Quotient Sets in Classification Problems. *Cybernetics and Systems Analysis*, 2014, vol. 50, iss. 4, pp. 507-518. DOI: 10.1007/s10559-014-9639-z.

22. Youssef, I. K., & Hassan, M. H. M. A Comparative Study for Some Mathematical Models of Epidemic Diseases with Application to Strategic Management. *Applied Sciences*, 2022, vol. 12, iss. 24, article no. 12639. DOI: 10.3390/app122412639.

23. Bazilevych, K., Butkevych, M., & Dotsenko, N. Cardiac Studies Diagnostic Data Informative Features Investigation based on Cumulative Frequency Analysis. *CEUR Workshop Proceedings*, 2022, vol. 3348, pp. 84–89.

24. Davidich, N., Chumachenko, I., Davidich, Y., Taisiia, H., Artsybasheva, N., & Tatiana, M. Advanced Traveller Information Systems to Optimizing Freight Driver Route Selection. *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*, 2020, pp. 111-115. DOI: 10.1109/dese51703.2020.9450763.

25. Zanini, A., & Woodbury, A. D. Contaminant source reconstruction by empirical Bayes and Akaike's Bayesian Information Criterion. *Journal of Contaminant Hydrology*, 2016, vol. 185-186, pp. 74–86. DOI: 10.1016/j.jconhyd.2016.01.006.

26. Krak, I., Kudin, H., Kasianiuk, V., & Efremov, M. Hyperplane Clustering of the Data in the Vector Space of Features Based on Pseudo Inversion Tools. *CEUR Workshop Proceedings*, 2021, vol. 3003, pp. 98-105.

27. Ying, J., Wang, Q., Xu, T., & Lu, Z. Diagnostic potential of a gradient boosting-based model for detecting pediatric sepsis. *Genomics*, 2021, vol. 113, iss. 1, pp. 874-883. DOI: 10.1016/j.ygeno.2020.10.018.

28. Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, 2020, vol. 121, article no. 103761. DOI: 10.1016/j.compbiomed.2020.103761.

29. Nizri, M. *COVID-19 Dataset*. 2023 [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/meirnizri/covid19-dataset (Accessed 23 May 2023).

ПОРІВНЯЛЬНИЙ АНАЛІЗ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ РІВНІВ РИЗИКУ ПАЦІЄНТІВ З COVID-19

Ксенія Базілевич, Олена Кириленко, Юрій Парфенюк, Сергій Кривцов,
Євген Меняйлов, Вікторія Кузнецова, Дмитро Чумаченко

Пандемія COVID-19 поставила перед світовими системами охорони здоров'я безпрецедентні виклики, акцентуючи необхідність прогнозних інструментів для розподілу ресурсів та догляду за пацієнтами. Це дослідження занурюється в потенціал моделей машинного навчання для прогнозування оцінки ризику пацієнтів з COVID-19, використовуючи всеосяжний набір даних. **Метою** цього дослідження було оцінити та порівняти ефективність трьох методологій машинного навчання – Байєсівського критерію, логістичної регресії та градієнтного бустінгу – у прогнозуванні ризику, пов'язаного з пацієнтами з COVID-19, на основі їх симптомів, статусу та медичної історії. **Об'єктом** дослідження є процес визначення стану пацієнта. **Предметом** дослідження є методи машинного навчання для визначення стану пацієнта. Для досягнення мети дослідження були сформульовані наступні **завдання**: слід проаналізувати методи та моделі визначення стану пацієнтів з COVID-19; слід розробити класифікаційну модель визначення стану пацієнта на основі критерію Байєса; слід розробити класифікаційну модель визначення стану пацієнта на основі логістичної регресії; слід

розробити класифікаційну модель визначення стану пацієнта на основі градієнтного бустінгу; слід розробити інформаційну систему; слід провести експериментальне дослідження на основі методів машинного навчання; слід проаналізувати результати експериментального дослідження. **Методи:** Використовуючи набір даних, наданий урядом Мексики, що охоплює понад мільйон унікальних пацієнтів з 21 відмінною ознакою, ми розробили інформаційну систему на мові програмування C#. Ця система дозволяє користувачам вибирати бажаний метод для розрахунку ризику, пропонуючи інструмент прийняття рішень у реальному часі для медичних фахівців. **Результати:** Усі моделі продемонстрували високі рівні точності. Однак були зауважені невеликі відмінності в їх показниках продуктивності, таких як чутливість, точність та показник F1. Метод градієнтного бустінгу трохи перевершив інші моделі за загальною точністю. **Висновки:** Хоча кожна модель продемонструвала свої переваги, вибір методу повинен базуватися на конкретних потребах та обмеженнях системи охорони здоров'я. Метод градієнтного бустінгу виявився дещо кращим у цьому дослідженні. Дослідження підкреслює потенціал машинного навчання у покращенні стратегій реагування на пандемію, пропонуючи як наукові інсайти, так і практичні інструменти для медичних фахівців.

Ключові слова: оцінка стану пацієнту; класифікація; машинне навчання; COVID-19; критерій Байєса; логістична регресія; градієнтний бустінг.

Базілевич Ксенія Олексіївна – канд. техн. наук, доц. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Кириленко Олена Юріївна – здобувач каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Парфенюк Юрій Леонідович – канд. техн. наук, викладач. каф. теоретичної та прикладної інформатики, Харківський національний університет ім. В. Н. Каразіна, Харків, Україна.

Кривцов Сергій Олегович – здобувач каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Меняйлов Євген Сергійович – канд. техн. наук, в.о. зав. каф. теоретичної та прикладної інформатики, Харківський національний університет ім. В. Н. Каразіна, Харків, Україна.

Кузнєцова Вікторія Олександрівна – канд. фіз.-мат. наук, доц. каф. вищої математики та інформатики факультету математики і інформатики, Харківський національний університет ім. В. Н. Каразіна, Харків, Україна.

Чумаченко Дмитро Ігорович – канд. техн. наук, доц., доц. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Kseniia Bazilevych – PhD in Information Technologies, Associate Professor of Mathematical Modelling and Artificial Intelligence Department, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: ksenia.bazilevich@gmail.com, ORCID: 0000-0001-5332-9545.

Olena Kyrylenko – Applicant of Mathematical Modelling and Artificial Intelligence Department, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: o.kyrylenko@khai.edu, ORCID: 0000-0001-5332-9545.

Yurii Parfenyuk – PhD in Automation and Computer-integrated Technologies, Lecturer of Theoretical and Applied Informatics Department, V. N. Karazin Kharkiv National University, Kharkiv, Ukraine,
e-mail: parfuriy.l@gmail.com, ORCID: 0000-0001-5357-1868.

Serhii Krivtsov – Applicant of Mathematical Modelling and Artificial Intelligence Department, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: krivtsovpro@gmail.com, ORCID: 0000-0001-5214-0927.

Ievgen Meniailov – PhD in Mathematical Modelling and Optimization Methods, Acting Head of Theoretical and Applied Informatics Department, V. N. Karazin Kharkiv National University, Kharkiv, Ukraine,
e-mail: evgenii.menyailov@gmail.com, ORCID: 0000-0002-9440-8378.

Victoriya Kuznietcova – PhD in Mathematical analysis, Associate Professor of Department of Higher Mathe-matics and Informatics, School of Mathematics and Computer Sciences, V. N. Karazin Kharkiv National University, Kharkiv, Ukraine,
e-mail: vkuznietcova@karazin.ua, ORCID: 0000-0003-3882-1333.

Dmytro Chumachenko – PhD in Artificial Intelligence, Associate Professor of Department of Mathematical Modelling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: dichumachenko@gmail.com, ORCID: 0000- 0003-2623-3294.