НОВІ МЕТОДИ І ТЕХНОЛОГІЇ

UDC 616.391:577.161.2-06:616-008.9]-07
DOI https://doi.org/10.32782/2226-2008-2024-5-12

A. V. Shanyhin*[1] *https://orcid.org/0000-0003-2644-4542
V. V. Babienko*[1] *https://orcid.org/0000-0002-4597-9908
A. M. Rozhnova*[1] *https://orcid.org/0000-0001-7718-6171
Ye. M. Strakhov*[2] *https://orcid.org/0000-0001-8207-8108
A. S. Korkhova*[2] *https://orcid.org/0009-0003-7234-5358

DEPENDENCE OF VITAMIN D LEVEL ON LABORATORY AND ANTHROPOMETRIC INDICATORS: APPLICATION OF MACHINE LEARNING METHODS FOR SCREENING IN ADULTS

[1]Odesa National Medical University, Odesa, Ukraine
[2]Odesa I. I. Mechnikov National University, Odesa, Ukraine

UDC 616.391:577.161.2-06:616-008.9]-07

V. V. Babienko[1], A. V. Shanyhin[1], A. M. Rozhnova[1], Ye. M. Strakhov[2], A. S. Korkhova[2]
DEPENDENCE OF VITAMIN D LEVEL ON LABORATORY AND ANTHROPOMETRIC INDICATORS: APPLICATION OF MACHINE LEARNING METHODS FOR SCREENING IN ADULTS
[1]Odesa National Medical University, Odesa, Ukraine
[2]Odesa I. I. Mechnikov National University, Odesa, Ukraine

Vitamin D deficiency is now recognized as an international health issue, affecting a variety of physiological systems and disease outcomes.

Purpose. The present study proposes machine learning models to identify individuals at risk of vitamin D deficiency.

Materials and methods. Machine learning was used on the dataset of 944 persons' laboratory analysis to determine the list of anthropometric and laboratory indicators that affect the development of vitamin D deficiency. It was built a decision tree with a depth of 5 to predict vitamin D deficiency based on various parameters.

Results. The authors found feature importance in identifying potential vitamin D deficiency. Age and BMI were considered the most impactful anthropometric parameters, level of HDL was the most important laboratory parameter. A heatmap matrix for correlation of features between one another was created. It was calculated metrics based on the confusion matrix for determining the risk of a 25(OH)D deficit: Accuracy, Precision, Sensitivity, Specificity, F1-Score. The authors plotted the ROC curve of the optimal model; established that the Area Under the Curve (AUC) of the selected model is equal to 0.92 that is a very effective result.

Conclusion. Machine learning techniques are more effective at predicting deficiencies than traditional statistical methods.

Key words: vitamin D, prevention, lipid metabolism, anthropometry, artificial intelligence, machine learning.

УДК 616.391:577.161.2-06:616-008.9]-07

А. В. Шанигін[1], В. В. Бабієнко[1], А. М. Рожнова[1], Є. М. Страхов[2], А. С. Корхова[2]
ЗАЛЕЖНІСТЬ РІВНЯ ВІТАМІНУ D ВІД ЛАБОРАТОРНИХ ТА АНТРОПОМЕТРИЧНИХ ПОКАЗНИКІВ: ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ З МЕТОЮ СКРИНІНГУ У ДОРОСЛИХ
[1]Одеський національний медичний університет, Одеса, Україна
[2]Одеський національний університет імені І. І. Мечникова, Одеса, Україна

Метою дослідження було визначення переліку антропометричних та лабораторних показників, що впливають на розвиток дефіциту вітаміну D, та розробка надійної прогностичної моделі, яка допоможе в ранньому виявленні та варіантах корекції дефіциту та недостатності вітаміну D у групах ризику. Машинне навчання було використано на наборі даних лабораторного аналізу 944 осіб, які впливають на розвиток дефіциту вітаміну D. З'ясовано важливість ознак у виявленні потенційного дефіциту вітаміну D. Вік та ІМТ вважалися найбільш впливовими антропометричними параметрами, рівень ЛПВЩ був найважливішим лабораторним параметром. Розраховані показники, що створені на основі матриці невідповідностей, для визначення ризику дефіциту 25(OH)D. Побудовано ROC-криву оптимальної моделі, яка доводить її ефективність.

Ключові слова: вітамін D, профілактика, ліпідний обмін, антропометрія, штучний інтелект, машинне навчання.

Introduction. Vitamin D deficiency is now recognized as an international health issue, affecting a variety of physiological systems and disease outcomes [1]. Despite its vital role in bone health, immunological function, and general well-being, vitamin D insufficiency is nonetheless common around the globe, impacting people of all ages and races [1; 2]. The level of vitamin D in the body is affected by a large number of different factors [3]. Among the risk factors for the development of deficiency and insufficiency of vitamin D, the most significant are insufficient exposure to the sun, a deficiency in the diet of products containing the daily level of vitamin D, dark skin color, excess body weight and obesity, disorders of digestive processes caused by malabsorption, Crohn's disease, gluten intolerance, kidney disease [1; 4; 5]. This emphasizes the need for novel approaches to predicting and preventing the development of vitamin D insufficiency and deficiency in the population.

Machine learning, an area of artificial intelligence, allows computers to learn from data in order to make predictions. The use of artificial intelligence in medicine has attracted the attention of scientists due to its potential to transform the system of forecasting, diagnosis and treatment of various pathological conditions. In recent years, more and more publications have appeared devoted to the use of machine learning algorithms for researching the risks of developing vitamin D deficiency and its correction [5; 6]. Machine learning models are capable of identifying patterns and relationships that might not be immediately obvious using conventional statistical methods through the use of large-scale datasets and advanced analytical tools.

The present study investigates the application of machine learning algorithms to predict vitamin D deficiency, drawing on a wide range of data sources such as demographic information, genetic markers, anthropometric and laboratory features.

Incorporating machine learning to predict vitamin D insufficiency offers enormous potential to improve knowledge of the complex etiology of vitamin D deficiency and insufficiency and inform timely, individualized preventive measures for each patient depending on the etiology of vitamin D deficiency. By creating accurate predictive models, healthcare professionals can identify patients who have a higher risk of vitamin D deficiency or insufficiency, and prescribe therapy in a timely manner [1; 5].

The implementation of machine learning in health care institutions will contribute to early diagnosis and prevention of vitamin D deficiency in risk groups [4; 5; 7]. The incorporation of an early screening system based on artificial intelligence will help reduce the burden on laboratories, and as a result, reduce costs for the health care system [1; 4; 8; 9].

Objective – to determine the list of anthropometric and laboratory indicators that affect the development of vitamin D deficiency; to develop a reliable prognostic model that will help in early detection and options for the correction of vitamin D deficiency and insufficiency in risk groups.

Materials and methods. In the course of the study, 944 persons aged from 20 to 91 years (average age – 46.9 years) were examined in order to create a dataset containing anthropometric and laboratory parameters from an assortment comprising people with and without vitamin D insufficiency. The variables in the dataset were sex, total cholesterol, level of high-density lipoproteins (HDL), level of low-density lipoproteins (LDL), level of very low-density lipoproteins (VLDL), atherogenicity coefficient (AC), triglycerides (TG), body mass index (BMI), waist circumference, waist-to-hip ratio (WHR), age, and vitamin D status (deficiency vs. sufficiency). The dataset was split into training and testing sets and decision tree methods were applied to develop predictive models. The evaluation of model performance was conducted using such metrics as accuracy, precision, sensitivity, specificity and the F1-score. In addition, feature significance analysis was used to determine the most important markers of vitamin D deficiency. The best tree was created, and the receiver operating characteristic curve and area under the curve (AUC) values were calculated.

Examination of patients was carried out on the basis of private medical centers of southern Ukraine "Yes Medical" and "Artromed". Further observation was carried out on an outpatient basis. All patients participating in the study were given oral and written information about the purpose and objectives of the study. Patients had the option to withdraw from the study at any time without giving a reason. Information about consent to participate in the study was documented by bilateral signing of the relevant document. The research was carried out with the provision of safety measures for life and health, with respect for human rights and moral and ethical standards, which corresponds to the principles of the Helsinki Declaration of Human Rights and the order of the Ministry of Health of Ukraine No. 693 dated 01.10.2015, the Council of Europe Convention on Human Rights and of biomedicine (ETS-164) dated 04.04.1997, the Status of the Ukrainian Association for Bioethics and GCP norms (1992) and approved by the commission on bioethics of Odessa National Medical University (protocol No. 12 dated 12.23.2019).

Results and Discussion. The previous study [5] used logistic regression to predict vitamin D deficiency through the use of demographic, clinical, and laboratory data. The logistic regression model used a binary classification framework to predict vitamin D deficiency based on factors such as age, gender, BMI, and laboratory results. While logistic regression provided useful insights into the relationship between variables and the probability of vitamin D deficiency, it was limited in its ability to capture nonlinear correlations and interactions among them.

In contrast, decision trees offer a flexible and intuitive method for representing complex data interactions. Decision trees divide the feature space into subsets based on simple decision rules, resulting in comprehensible decision paths. By recursively partitioning the data based on the most significant features, decision trees can capture complex decision boundaries and interactions among predictors.

Figure 1 depicts the visualization of the top levels of the optimal decision tree model to predict vitamin D deficiency based on various demographic, clinical, and laboratory parameters. With a maximum depth of 5, the decision tree can generate up to 5 levels of splits, each reflecting a decision rule based on a distinct feature.

In particular, vitamin D deficiency can occur under the following conditions: age up to 28.5 years, body mass index

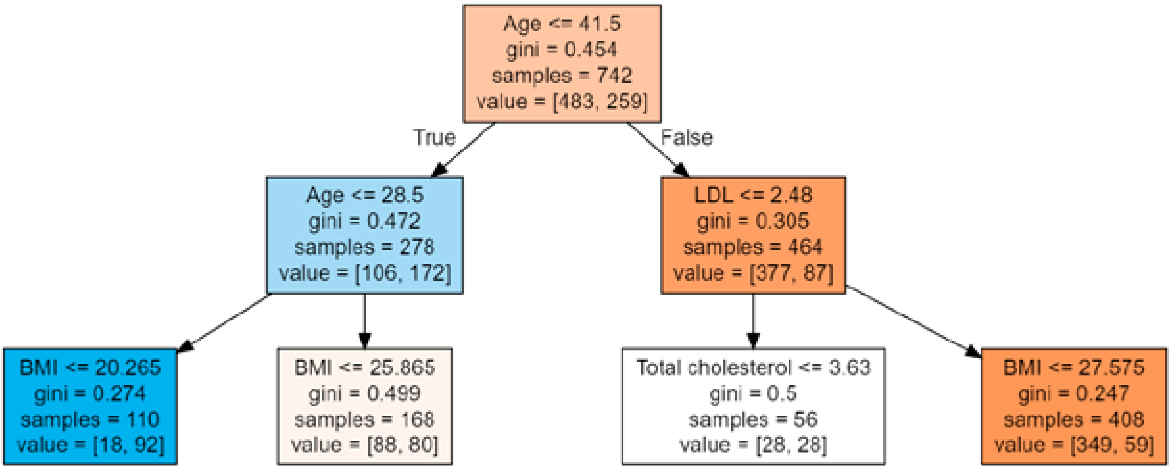

Fig. 1. Visualization of the three top levels of the decision tree model

above 20.265 kg/m^2, total cholesterol above 3.3 mmol/l, high-density lipoprotein level below 2.21 mmol/l.

Figure 2 displays the results of the decision tree method's analysis of feature importance in identifying potential vitamin D deficiency.

The results of the feature importance analysis show that anthropometric parameters such as age and BMI are the most impactful. The level of HDL is the most important laboratory parameter. When correlation values between features exceed 0.5, they are considered dependent. The detailed heatmap matrix for feature correlations is shown in Figure 3.

A confusion matrix, sometimes referred to as an error matrix, is a table used in machine learning to illustrate how effectively a classification method performs (Table 1). The actual values of the classes are shown in the rows of this matrix, whereas the predicted classes are shown in the columns. There are four potential categories of outcomes: true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) [10].

The following metrics are created based on the confusion matrix:

Accuracy=(*TP*+*TN*)/(*TP*+*TN*+*FP*+*FN*) – indicates how frequently the classifier is true;

Precision=*TP*/(*TP*+*FP*) – indicates the degree to which one may "believe" that the model would predict class 1, i.e., that the item is a member of the 25(OH)D level deficient class;

Sensitivity=*TP*/(*TP*+*FN*) – indicates how effectively it identifies positive class 1;

Specificity=*TN*/(*TN*+*FP*) – indicates how well it can detect the negative class 0;

F1 Score=(2·Precision·Sensitivity)/(Precision+Sensitivity) – the harmonic mean of Precision and Sensitivity.

Metrics for determining the risk of a 25(OH)D deficit are shown in Table 2.

Additionally, the models were assessed using Receiver Operating Characteristic (ROC) analysis employing graphs and ROC curves. The link between the sensitivity and specificity of the model is depicted by the ROC curve, which displays the dependence of the number of correctly categorized positive examples on the number of erroneously classified negative cases. Figure 4 shows the ROC curve of the optimal model.

As an examination of the ROC curve, the Area Under the Curve (AUC) represents a binary classifier's ability to differentiate between classes. The model performs better

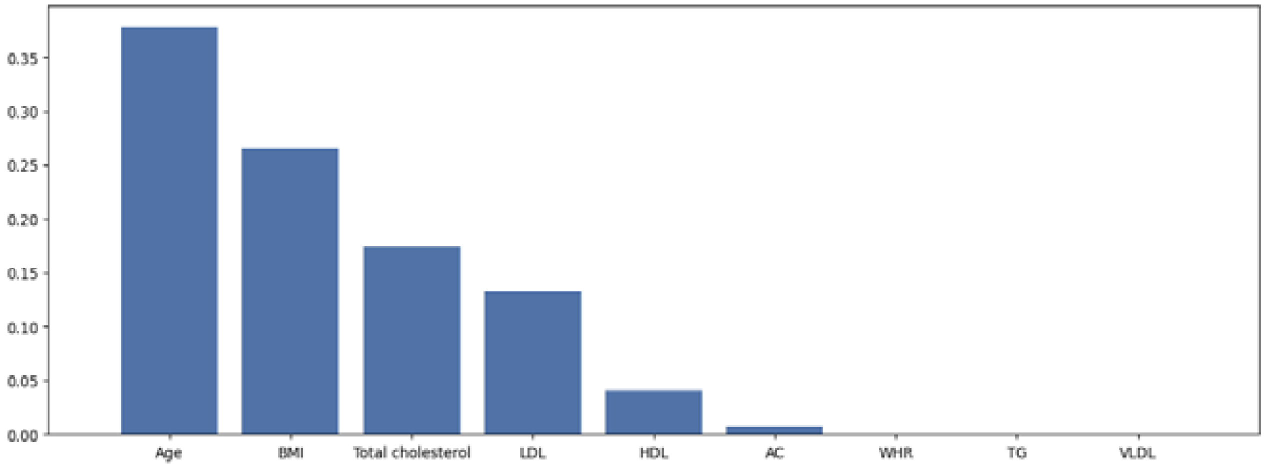

Fig. 2. Feature importance

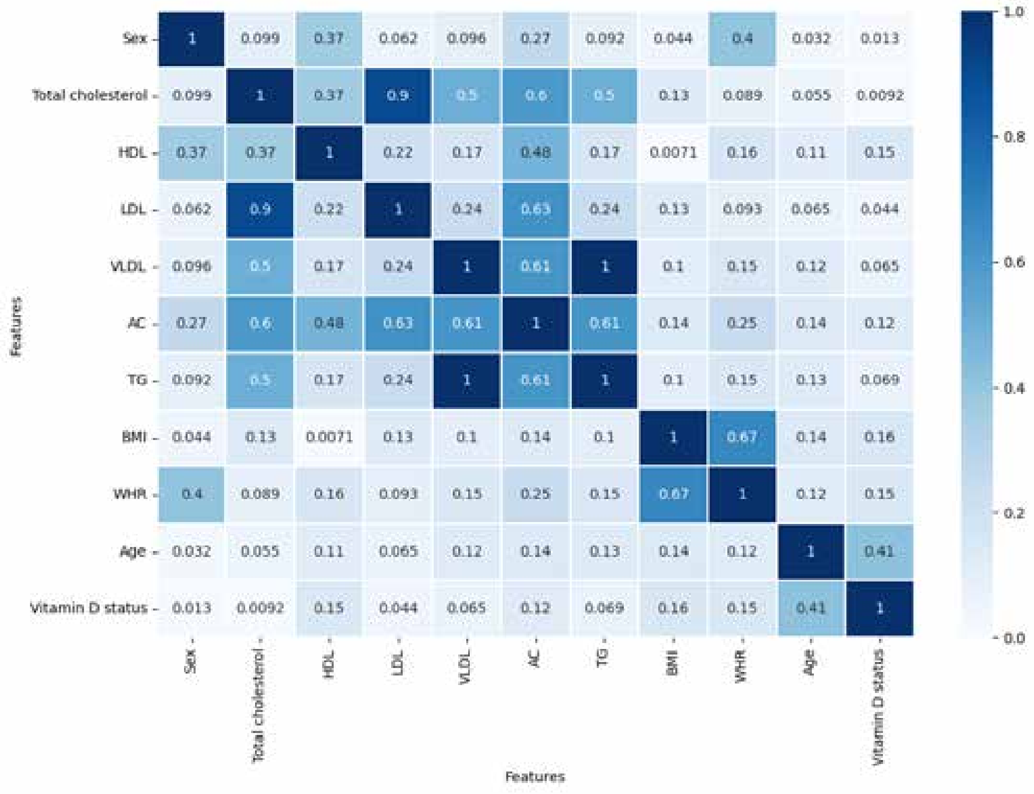

Fig. 3. Feature importance using correlation matrix heatmap

Table 1

Confusion matrix based on classification results

		Predicted	
		Negative (0)	**Positive (1)**
Actual	Negative (0)	TN	FP
	Positive (1)	FN	TP

Table 2

Evaluation metrics for assessment of 25(OH)D deficiency risk

Evaluation metric	**Value obtained on the train sample**	**Value evaluated on the test sample**
Accuracy	0.914	0.913
Precision	0.901	0.932
Sensitivity/Recall	0.859	0.821
Specificity	0.946	0.966
F1-Score	0.879	0.873

at discerning between the positive and negative classes the higher the AUC. The classifier has a good probability of differentiating between positive and negative class values when 0.5<AUC<1. The reason for this is that the classifier can identify a higher proportion of True positives and True negatives than False positives and False negatives. The AUC of the selected model is 0.92, indicating high accuracy of the model, as it exceeds 0.9. In comparison, AUC values in the range between 0.80 and 0.90 are considered to be sufficiently accurate, and values between 0.70 and 0.80 are considered poor.

Conclusions. The following metrics were used to evaluate the effectiveness of machine learning methods: Accuracy, Precision, Recall, Specificity, and F1-score. The study shows that decision tree models are more reliable for forecasting deficits than traditional statistical methods or linear models, particularly logistic regression. This study improved performance in terms of forecasting accuracy compared to previously published regression models. Machine learning models based on decision trees capable of identifying individuals at risk of vitamin D deficiency have been built and validated using comprehensive datasets that include demographic, anthropometric, and laboratory information. The findings highlight machine learning's

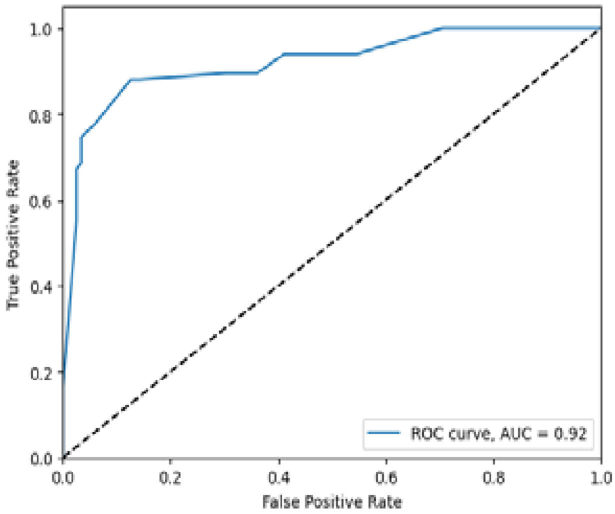

Fig. 4. Receiver operating characteristic

potential to improve preventative medicine through focused interventions and tailored healthcare approaches.

The study's key highlights are:

1. Increased prediction accuracy: Decision tree algorithms outperformed logistic regression in predicting vitamin D deficiency. The use of nonlinear algorithms allowed for the detection of complex correlations and patterns, resulting in higher accuracy and AUC-ROC values.

2. Identification of Key Predictors: Feature importance analysis identified age, body mass index, and HDL level as significant predictors of vitamin D deficiency. These findings can help healthcare practitioners develop effective risk assessment and intervention strategies.

3. Clinical implications: The use of machine learning models in clinical practice shows potential for improving patient outcomes and reducing the burden of vitamin D-related disorders. By correctly identifying patients at risk of deficiency, healthcare providers can apply specific treatments, such as supplementation, dietary changes, and lifestyle guidance, to improve vitamin D status and prevent health consequences.

4. Future Directions: While the study represents a significant step forward, further research is needed to address several critical aspects. Prospective validation of the models in diverse populations and clinical settings is crucial for determining generalizability and robustness. Additionally, recent advancements in machine learning techniques, such as deep learning and ensemble methods, may further improve prediction accuracy and model interpretability.

Ultimately, the study emphasizes the potential of machine learning in predicting vitamin D insufficiency and developing customized treatment strategies. By leveraging data-driven techniques, healthcare professionals can gain the insights and tools necessary to enhance patient health and well-being.

BIBLIOGRAPHY

1. Grygorieva NV, Tronko MD, Kovalenko VM, et al. Diagnosis, prevention and treatment of vitamin D deficiency in adults: Ukrainian experts consensus statement. *PAIN, JOINTS, SPINE.* 2023; 13(2): 60–76. doi: https://doi.org/10.22141/pjs.13.2.2023.368 (in Ukrainian).
2. Abboud M, Liu X, Fayet-Moore F, et al. Effects of Vitamin D Status and Supplements on Anthropometric and Biochemical Indices in a Clinical Setting: A Retrospective Study. *Nutrients.* 2019;11(12):3032. doi: https://doi.org/10.3390/nu11123032.
3. Arredondo A, Azar A, Recamán AL. Diabetes, a global public health challenge with a high epidemiological and economic burden on health systems in Latin America. *Glob. Public Health.* 2018; 13(7): 780–787. doi: https://doi.org/10.1080/17441692.2017.1316414.
4. Shanygin AV. The significance of diet and insolation levels in vitamin D supply. Modern aspects of prevention. *Health of Society*. 2022;11(1):16–22. doi: https://doi.org/10.22141/2306-2436.11.1.2022.288 (in Ukrainian).
5. Shanyhin A, Babienko V, Strakhov Ye, Korkhova A. Mathematical modeling of the dependence of the risk of vitamin D deficiency on anthropometric and laboratory parameters. *Journal of Education, Health and Sport.* Online. 2023; 13 (4): 356–366. [Accessed 28 February 2024]. doi: https://doi.org/10.12775/JEHS.2023.13.04.042.
6. Netrebin L, Pankiv V, Kyryliuk M. Mathematical model for assessing the prognostic significance of 25(OH)D deficiency in the progression of diabetic retinopathy in type 2 diabetes patients. *Mìžnarodnij endokrinologìčnij žurnal* [Internet]. 2023 [cited 2024 May 30]; 19(4): 269–73. Available from: https://iej.zaslavsky.com.ua/index.php/journal/article/view/1284 (in Ukrainian).
7. Calame W, Street L, Hulshof T. Vitamin D Serum Levels in the UK Population, including a Mathematical Approach to Evaluate the Impact of Vitamin D Fortified Ready-to-Eat Breakfast Cereals: Application of the NDNS Database. *Nutrients*. 2020; 12(6): 1868. doi: https://doi.org/10.3390/nu12061868.
8. Setayesh L, Amini A, Bagheri R, et al. Elevated Plasma Concentrations of Vitamin D-Binding Protein Are Associated with Lower High-Density Lipoprotein and Higher Fat Mass Index in Overweight and Obese Women. *Nutrients*. 2021; 13(9): 3223. doi: https://doi.org/10.3390/nu13093223.
9. Sizar O, Khare S, Goyal A, et al. Vitamin D Deficiency. [Updated 2023 Jul 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK532266/.
10. Ting KM. Confusion Matrix. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. 2011. Available from: https://doi.org/10.1007/978-0-387-30164-8_652.

Надійшла до редакції 20.06.2024 р.
Прийнята до друку 26.12.2024 р.
Електронна адреса для листування anton.shanyhin@onmedu.edu.ua