ACTA METALLURGICA SLOVACA
2022, VOL. 28, NO. 2, 91-96

RESEARCH PAPER

AN APPROACH TOWARD PREDICTION OF SM-CO ALLOY'S MAXIMUM ENERGY PRODUCT USING FEATURE BAGGING TECHNIQUE

Andrii Trostianchyn[1], Zoia Duriagina[1,2], Ivan Izonin[3], Roman Tkachenko[3], Volodymyr Kulyk[1], Natalia Lotoshynska [3]*

[1]*Department of Materials Science and Engineering, Lviv Polytechnic National University, Lviv, 79013, Ukraine*
[2]*The John Paul II Catholic University of Lublin, Al. Racławickie 14, 20-950 Lublin, Poland*
[3] *Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, 79013, Ukraine*
[4] *Department of Publishing Information Technologies, Lviv Polytechnic National University, Lviv, 79013, Ukraine*

*Corresponding author: kulykvolodymyrvolodymyrovych@gmail.com, tel.: +380672823572, *Department of Materials Science and Engineering, Lviv Polytechnic National University, Lviv, 79013, Ukraine*

Received: 05.04.2022
Accepted: 26.04.2022

ABSTRACT

This paper aims to solve the predicting magnetic properties task on the example of Sm-Co alloy using machine learning tools. In particular, the authors solved the Sm-Co alloys maximum energy product prediction task using the feature bagging technique. To implement this approach, we have chosen the Random Forest algorithm, which efficiently processes short datasets by reducing variance and, as a result, reducing the impact/avoidance of overfitting. Experimental modeling was based on a short set of data (190 observations) collected by the authors with many independent attributes. As a result, it has been experimentally established that the studied machine learning method provides a high value of the Coefficient of determination (0.78) when solving Sm-Co alloy's maximum energy product prediction task. Furthermore, by comparing with other ensemble-based methods from different classes, the highest accuracy of the researched process is established based on several performance indicators.

Keywords: Computational Material Science, machine learning, prediction model, small data processing, Sm-Co alloy, magnetic properties.

INTRODUCTION

A new approach to solving applied tasks in Material Science, namely Computational Material Science, is being intensively developed [1-3]. The use of various methods of computer modeling can be significantly shorter than the traditional methods for investigation. Also, it can reduce the cost and simplify both the process of creating new functional materials and improving the properties of existing ones [4, 5]. Furthermore, a vast arsenal of machine learning tools allows you to solve clustering, regression, and classification tasks successfully, predict the properties of materials, reveal hidden relationships, etc. [6, 7].

The well-known relationship between the composition (chemical, phase) of the material, microstructure, and different properties allows us to consider each experimental observation (e.g., measurement or calculation of a property taking into account processing parameters, composition, and structure) as a data point (vector) to create a database. Based on it, it becomes possible to build predictive models, which, among other things, can be used to predict the properties of materials. In addition to significantly accelerating and reducing the cost of experimental research, the construction of such models allows you to create machine learning models that are much more complex and critical to discovering and developing new materials. An example of such models is the prediction of fatigue characteristics of steel based on a set of experimental data [8], predicting the stability of compounds using modeling based on the theory of density functional (DFT method). On the other hand, the discovery of stable ternary compounds [9] and the optimization of structural parameters to improve the properties of magnetoelastic materials [10] was carried out using other models [2, 11].

Particularly relevant is the use of machine learning tools for processing Big Data, characterized by nonlinear, complex, and often unknown relationships between many variables [12]. However, the actual existence of preliminary, usually experimentally established data is the main factor that determines the possibility of using machine learning methods in Materials Science [13].

One of the examples is predicting the magnetic properties of permanent magnets based on rare earth metals (REM). In this case, it is necessary to take into account the chemical and phase composition of the material, crystallographic features, microstructure parameters, the size of the structural components, and so on. Our literature review revealed that in the case of Sm-Co alloys, there is a relatively limited number of publications that would contain all the information necessary to create an initial database. On the other hand, it was found that in almost all considered publications, the magnetic properties are represented by the coercive force H_c. At the same time, data on the saturation magnetization M_s, remanence $M_{r,}$ and maximum energy product $(BH)_{max}$ are presented to a lesser extent [14]. It is worth noting that minor information regarding the values $(BH)_{max}$ is available.

It is known that the possibility of using machine learning tools to solve specific Material Science tasks depends on both the number of available observations and the quality of processing of data selected for the dataset [13]. Therefore, in addition to past data availability, it is crucial to pre-process them to ensure acceptable quality – monitoring and removal of anomalies and

DOI: 10.36547/ams.28.2.1462

outliers in data, data deduplication, missing data recovery, etc. As a result, the number of vectors in the database significantly reduces, often making it impossible to create an adequate model. The main stages of data pre-processing include data sampling, missing data recovery, anomaly detection, normalization, attribute type conversion, feature selection, etc. Next, various supervised data analysis methods are used for predictive modeling.

Based on this, to assess the real possibility of predicting the magnetic properties of ferromagnetic alloys of the Sm-Co system, at the first stage, we attempted to use machine learning methods to predict coercive force based on the collected dataset [14]. In particular, an experimental comparison of eight existing machine learning methods was conducted to select the optimal techniques for building a stacking ensemble model based on heterogeneous elements. As a result, there is a significant increase in the accuracy of the proposed model compared to single-based algorithms that formed it (Neural Networks, AdaBoost, Gradient Boosting, and Random Forest algorithm) and other machine learning methods (SVR, SGD, Linear regression, and Tree). The high prediction accuracy of the proposed stacking model makes it possible to use it to predict the coercive force of Sm-Co alloys.

However, such a strategy requires considerable energy and computational costs to implement each individual method from the stacking ensemble. In addition, the disadvantage of this approach is the need to adjust a large number of different parameters of all the methods underlying the work of stacking.

However, a significant reduction in the number of vectors that can be used to predict the maximum energy product requires testing the possibility of using ensemble-based strategies of other classes to predict this property based on the collected data set. Thus, solving the Sm-Co alloy's maximum energy product prediction task, particularly with machine learning tools, is a topical task. It should be noted that the current development of Computational Materials Science encourages the use of such a strategy in various application areas, which are characterized by a limited amount of available data. However, such approaches to solving the stated task are insufficiently covered, particularly in the scientific literature. At the same time, the successful solution of the task can significantly contribute to the successful completion of many studies aimed at optimizing the properties of functional materials.

In the case of the ferromagnetic materials based on Sm-Co alloys, we are talking about a significant reduction in time, financial and other costs associated with the production of prototypes of permanent magnets, and the study of their magnetic properties. The fact is that the creation of a new generation of permanent magnets based on REM involves the development of new technological approaches to obtaining magnetoanisotropic powders of such alloys in the nanostructured state [15]. We used mechanothermal treatment in hydrogen to get such powders [16-18]. It is shown that by changing the processing conditions, it is possible to control the phase composition of materials and influence the features of the microstructure. It is essential to establish the influence of these parameters on the studied materials' magnetic properties to select the optimal modes of hydrogen treatment. That is why this paper aims to accurately predict Sm-Co alloy's maximum energy product via machine learning method in the case of a limited amount of data.

The novelty of the results obtained in the article is as follows:

- we have collected the data set, pre-processed it, and applied the feature bagging technique to solve the prediction of Sm-Co alloy's maximum energy product task;
- we have used Random Forest as a high-precision algorithm that implements a feature bagging strategy; we have conducted experimental modeling and selected the optimal values of the parameters of its work;
- we have established the highest accuracy of the studied machine learning method compared to other ensemble-based methods using four different performance indicators.

MATERIAL AND METHODS

Dataset description

The dataset is based on the processing and analysis of a large amount of literature data on the dependence of the magnetic properties of Sm-Co alloys on their chemical composition, phase composition, state of the material, presence or absence of crystallographic texture of the main ferromagnetic phase and microstructure parameters, including structural components (Fig. 1). As a result, we created a dataset containing 419 observations, each of which is described by 31 variables. It should be noted that the value of Hc as a target value is available for all observations, while the value of Mr for 411 and (BH) max for only 190 vectors. A detailed description of the process of creating a dataset, its features, adopted simplicity, and a list of references used is given in [14]. The main difference between the dataset used in this paper and its full version is that it does not contain vectors for which there are no maximum energy product (BH) max values as the target value. In addition, input data do not include a variable describing the total content of 3d transition metals (Ti and Ni). The reason was the lack of information about their presence for all observations used in this dataset.

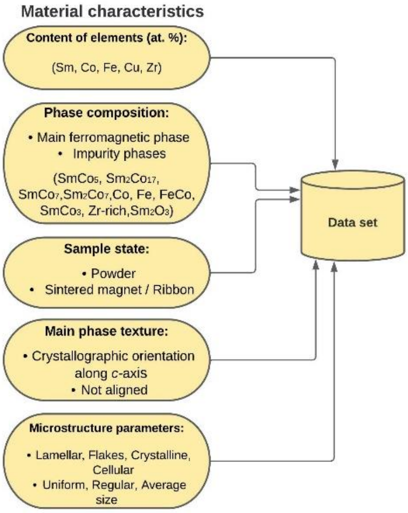
Material characteristics
Content of elements (at. %):
(Sm, Co, Fe, Cu, Zr)
Phase composition:
• Main ferromagnetic phase
• Impurity phases
($SmCo_5$, Sm_2Co_{17}, $SmCo_7$,Sm_2Co_7,Co, Fe, FeCo, $SmCo_3$, Zr-rich,Sm_2O_3)
Sample state:
• Powder
• Sintered magnet / Ribbon
Main phase texture:
• Crystallographic orientation along c-axis
• Not aligned
Microstructure parameters:
• Lamellar, Flakes, Crystalline, Cellular
• Uniform, Regular, Average size
Data set

Fig. 1 The attributes of the collected datasets

Thus, the collected dataset for solving the task of predicting the maximum energy product of ferromagnetic alloys of the Sm-Co system contains 190 observations, each of which is described by 30 variables. The dataset is available online at [19].

General methodology

The research methodology involves collecting data by experts, their preliminary processing and preparation, selection and justification of the machine learning model, modelling and evaluation of results (Fig. 2). Because experts collected the data for solving Sm-Co alloy's maximum energy product prediction task, feature selection procedures were not performed in this paper.

This is because each attribute of the collected data set significantly affects the result of the prediction of the magnetic properties of the alloy.

The authors have collected a set of data containing 190 observations. Collecting more data is a resource-intensive, time-consuming, and materially costly procedure. Therefore, the paper aimed to intellectually process a short dataset to obtain accurate prediction results, sufficient for using the studied model in practice.

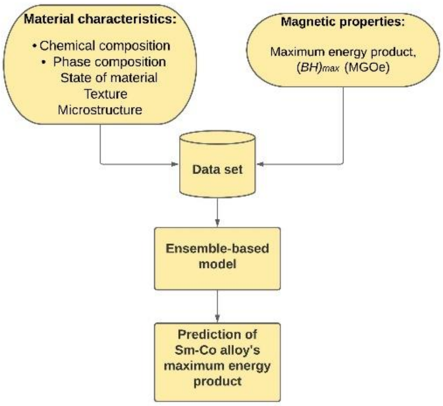

Fig. 2 General methodology

The collected dataset is significantly nonlinear, so linear models will not provide sufficient prediction accuracy. Moreover, most single-based models and Artificial Neural Networks are characterized by overfitting when processing a short dataset. Nevertheless, the chosen model should:

- Efficiently process short datasets with a large number of independent attributes that are characterized by complex interconnections.
- Ensure maximum prediction accuracy.
- Ensure a minimum duration of the training procedure.

That is why the authors chose the strategy of ensemble learning to solve the prediction of Sm-Co alloy's maximum energy product task. In addition to reducing the above disadvantages, this approach will minimize variance and avoid overfitting.

Feature bagging technique

The strategy of assembling machine learning methods involves using several weak classifiers or regressors to build a strong one. The primary aim of this step is to increase the prediction or classification accuracy.

In the literature, there are several fields of development of ensemble methods. The main ones are: boosting, bagging, and stacking. Each of them has its advantages and disadvantages, but we focused on the improved Bootstrap Aggregation or Bagging [20].

The basic version of the ensemble based on the bagging strategy involves using several algorithms (mostly decision trees) on small parts of the sample and generalizing the result. This approach reduces variance and, as a result, reduces the overfitting effect that is typical for the processing of short datasets. However, the main problem with this approach is that it uses a greedy algorithm for a variable split. Thus, the decision trees of such an ensemble can be very structurally similar and have a high correlation in their predictions.

Therefore, we used the feature bagging technique in this work, which was successfully implemented in the Random Forest algorithm [21]. This approach is based on a straightforward random sampling procedure (instead of choosing the most optimal split-point, as in basic bagging), which provides predictions from different sub-trees that are weakly correlated. That is why it ensures the success of the studied algorithm in solving regression and classification tasks in various application areas. Therefore, we will choose this model for our research.

RESULTS AND DISCUSSION

Modelling and results

The modeling process of the studied machine learning method to solve the prediction of Sm-Co alloy's maximum energy product task took place using the environment of intellectual analysis and data visualization - Orange software [22]. This choice is due to a user-friendly and intuitive graphical interface, a wide range of machine learning models, and many different widgets for visualization.

The modeling took place on a computer DELL, Intel CORE I7, 8Gb. Among the Random Forest algorithm parameters established experimentally, which provide the highest accuracy are: trees number – 10000, the subset that doesn't slip is equal to 5.

Among the performance indicators, we have used the following [23]:

Coefficient of determination (R2):

$$R^2 = 1 - \frac{\sum_{i=1}^{T}\left(y_i^{actual} - y_i^{new}\right)^2}{\sum_{i=1}^{T}\left(y_i^{actual} - \overline{y}_i\right)^2}, \quad (1.)$$

Mean Square Error (MSE):

$$MSE = \frac{1}{T}\sum_{i=1}^{T}\left(y_i^{actual} - y_i^{new}\right)^2, \quad (2.)$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{T}\sum_{i=1}^{T}\left|y_i^{actual} - y_i^{new}\right|, \quad (3.)$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{T}\frac{\left(y_i^{actual} - y_i^{new}\right)^2}{T}}, \quad (4.)$$

where: y_i^{actual} - actual value,

y_i^{new} - new value, obtained after prediction,

T - number of vectors,

$\overline{y}_i = \frac{1}{T}\sum_{i=1}^{T} y_i^{actual}$.

Taking into account all performance indicators (1)-(4) provides a complete assessment of the effectiveness of the studied methods.
The simulation results for 10-folds cross-validation [24] are summarised in **Table 1**.

Table 1 Values of the error's indicators for the method investigated

Performance indicators	Random Forest algorithm
MSE	18,963
RMSE	4,355
MAE	3,029
R2	0,787

As can be seen from the results shown in **Table 1**, the application of the feature bagging technique implemented through the Random Forest algorithm to solve the Sm-Co alloy's maximum energy product prediction task demonstrates promising results. In particular, the adequacy of the studied model in the conditions of processing a short dataset (190 observations) [19] is confirmed by the high value of the Coefficient of determination. In addition, overfitting, which is typical during the processing of short datasets, is not observed. This is due to the features of the Random Forest algorithm, which can efficiently process short datasets, reduce variance and therefore is not prone to overfitting [25, 26].
To demonstrate the effectiveness of the researched algorithm to solve the Sm-Co alloy's maximum energy product prediction task, we will compare its work with the machine learning methods of other class.

Comparison and discussion

Several other ensemble-based methods of the same or other classes were chosen to compare the work of the studied method. In particular, we have investigated the efficiency of the following methods:

- Gradient Boosting [27];
- CatBoost [28]
- AdaBoost [29]

Simulation of the existing machine learning methods from the ensemble class for comparison with the studied methods took place using Orange software [22]. The flowchart of the modeling process in the selected environment is shown in **Fig. 3**.

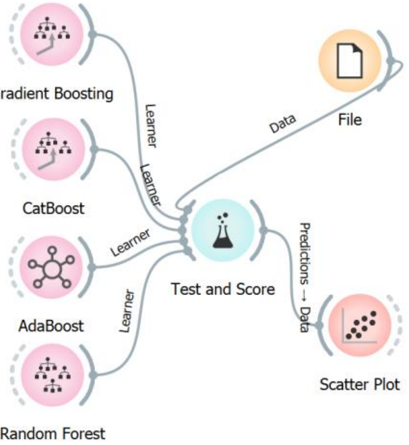

Fig. 3 Flowchart of the modelling process for different ensemble-based methods in Orange

The optimal parameter for all investigated methods was the following:

- Gradient Boosting regressor: number of trees is 1000; learning rate – 0.01; limit the depth of individual trees is 3; subset biggest that 2, the fraction of training instances is 1.0;
- CatBoost regressor: number of trees is 10000; learning rate is 1.0, regularization - lambda is 3; limit the depth of individual trees is 5, the fraction of training instances is 1.0;
- AdaBoost regressor: number of estimators is 10000; learning rate is 1.0, exponential regression loss function

10-folds cross-validation was also used to obtain reliable results. As a result, the following error values were obtained for all studied methods. It has been summarized in **Fig. 4**.

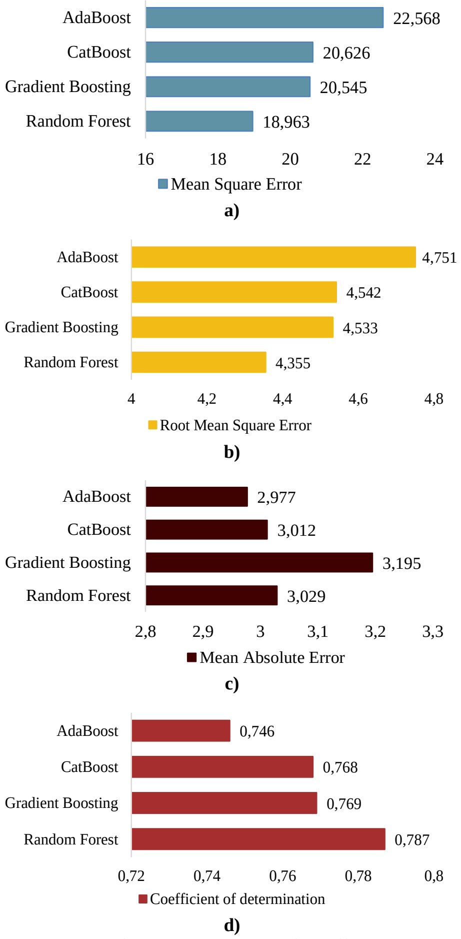

Fig. 4 Values of the error indicators for different ensemble-based methods: a) MSE; b) RMSE; c) MAE; d) R2

As can be seen from **Fig. 4**, two machine learning boosting methods, namely Adaboost and CatBoost, provide the lowest MAE error value (Fig. 4c). However, the boosting strategy shows the lowest results among all the considered methods in all other performance indicators. In particular, despite the fact that the models based on it are adequate (Fig. 4d) according to the value of the Coefficient of determination (R^2> 0.5), Random Forest here shows the highest value.

If we consider the values of all other metrics except MAE (Fig. 4 a), b), and d)), the studied machine learning algorithm provides the highest accuracy in solving the prediction of Sm-Co alloy's maximum energy product task. This result is due to the size of the data (small volumes) and the essential features of the Random Forest method. It makes it possible to use this method as a basis for an intelligent subsystem for predicting the magnetic properties of large applied systems in Computational Materials Science.

CONCLUSIONS

The paper considers the task of predicting magnetic properties on the example of Sm-Co alloy. The authors have applied a machine learning approach to solve it. In particular, the authors solved the predicting Sm-Co alloy's maximum energy product task using the feature bagging technique. This ensemble strategy is implemented in the form of the existing machine learning method - the Random Forest algorithm. That is why it was chosen to implement the modeling of the proposed approach.

The authors have collected a dataset containing 190 observations. Based on this, experimental modeling of the method was performed, and the optimal values of its operation parameters were selected. As a result, it has been experimentally established that the studied approach provides a high value of the Coefficient of determination - 0.78 when solving the predicting Sm-Co alloy's maximum energy product task.

By comparing with other ensemble-based methods of different classes, the highest accuracy of the studied approach is established based on various performance indicators.

Further research will increase prediction accuracy when solving the stated task. In particular, additional investigation will be conducted to build stacking models based on a set of different Artificial Intelligence tools. In addition, it is planned to develop hybrid variants of Artificial Neural Networks [30-32] in the particular General Regression Neural Network and homogeneous ensembles based on it [33, 34], which demonstrate high efficiency in processing short datasets.

Acknowledgments: The authors would like to thank the reviewers for the correct and concise recommendations that helped present the materials better. This work has become possible only because of the resilience and courage of the Armed Forces of Ukraine**.** The National Research Foundation of Ukraine funded this research under project number 2021.02/0041.

REFERENCES

1. Y. Hong, B. Hou, B. Jiang, J. Zhang: Wiley Interdisciplinary Reviews: Computational Molecular Science, 10(31), 2020, e1450. https://doi.org/10.1002/wcms.1450.
2. A. Agrawal, A. Choudhary: APL Materials, 4(5), 2016, 053208. https://doi.org/10.1063/1.4946894
3. J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques: npj Computational Materials, 5(1), 2019, 83. https://doi.org/10.1038/s41524-019-0221-0.
4. Y. Juan, Y. Dai, Y. Yang, J. Zhang: Journal of Materials Science and Technology, 79, 2021, 178-190. https://doi.org/10.1016/j.jmst.2020.12.010.
5. S. Ping Ong: Computational Materials Science, 161, 2019, 143-150. https://doi.org/10.1016/j.commatsci.2019.01.013.
6. L. Wang: Physical Review B, 94(19), 2016, 195105. https://doi.org/10.1103/PhysRevB.94.195105.
7. A. Jain, G. Hautier, S. P. Ong, K. Persson: Journal of Materials Research, 31(8), 2016, 977-994. https://doi.org/10.1557/jmr.2016.80.
8. A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, S. R. Kalidindi: Integrating Materials and Manufacturing Innovation, 3(1), 2014, 90-108. https://doi.org/10.1186/2193-9772-3-8.
9. B. Meredig et al.: Physical Review B - Condensed Matter and Materials Physics, 89(9), 2014, 094104. https://doi.org/10.1103/PhysRevB.89.094104.
10. R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, A. Choudhary: Scientific Reports, 5(23), 2015, 11551. https://doi.org/10.1038/srep11551.
11. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim: npj Computational Materials, 3(1), 2017, 54. https://doi.org/10.1038/s41524-017-0056-5.
12. L. Himanen, A. Geurts, A.S. Foster, P. Rinke: Advanced Science, 6(21), 2019, 1900808. https://doi.org/10.1002/advs.201900808.
13. Y. Liu, T. Zhao, W. Ju, S. Shi, S. Shi, S. Shi: Journal of Materiomics, 3(3), 2017, 159-177. https://doi.org/10.1016/j.jmat.2017.08.002.
14. A. Trostianchyn, Z. Duriagina, I. Izonin, R. Tkachenko, V. Kulyk, O. Pavliuk: Acta Metallurgica Slovaca, 27(4), 2021. 195-202. https://doi.org/10.36547/ams.27.4.1173.
15. O. Gutfleisch et al.: Advanced Materials, 23(7), 2011, 821-842. https://doi.org/10.1002/adma.201002180.
16. I. I. Bulyk, A. M. Trostyanchyn: Fiziko-Khimicheskaya Mekhanika Materialov, 39(4), 2003, 77-83.
17. I.I. Bulyk I.I., A.M. Trostyanchyn, P.Ya. Lyutyi: Materials Science, 48(3), 2012, 316-322. https://doi.org/10.1007/s11003-012-9508-8.
18. I.I. Bulyk: Materials Science, 54(6), 2019, 761-775. https://doi.org/10.1007/s11003-019-00262-7.
19. A. Trostianchyn, Z. Duriagina, I. Izonin, R. Tkachenko, V. Kulyk, N. Lotoshynska: Dataset for Sm-Co alloy's maximum energy product prediction task, 2022. https://www.researchgate.net/publication/359485980.
20. M.A. Yaman, F. Rattay, A. Subasi: Procedia Computer Science, 194, 2021, 202-209. https://doi.org/10.1016/j.procs.2021.10.074.
21. K. Fawagreh, M.M. Gaber, E. Elyan: Systems Science & Control Engineering: An Open Access Journal, 2(1), 2014, 602-609. https://doi.org/10.1080/21642583.2014.956265.
22. J. Demšar et al.: Journal of Machine Learning Research, 14, 2013, 2349-2353.
23. B. Rawat, S.K. Dwived: International Journal of Information Technology and Computer Science, 11(1), 2019, 14-23. https://doi.org/10.5815/ijitcs.2019.01.02.
24. I.K. Nti, O. Nyarko-Boateng, J. Aning: International Journal of Information Technology and Computer Science, 13(6), 2021, 61-71. https://doi.org/10.5815/ijitcs.2021.06.05.
25. A. Iqbal, S. Aftab, I. Ullah, M. S. Bashir, M. A. Saeed: International Journal of Modern Education and Computer Science, 11(9), 2019, 54-64. https://doi.org/10.5815/ijmecs.2019.09.06.
26. O.S.S. Alsharif, K.M. Elbayoudi, A.A.S. Aldrawi, K. Akyol: International Journal of Information Engineering and Electronic Business, 11(5), 2019, 19-23. https://doi.org/10.5815/ijieeb.2019.05.03.
27. P. Bhowmik, P.C. Bhowmik, U. A. Md. Ehsan Ali, Md. Sohrawordi: International Journal of Information Technology and Computer Science, 13(5), 2021, 30-40. https://doi.org/10.5815/ijitcs.2021.05.03.
28. F. Zhou et al.: Mathematical Problems in Engineering, 2021, 2021, 1929137. https://doi.org/10.1155/2021/1929137.
29. M. Z. Khan: International Journal of Modern Education and Computer Science, 12(1), 2020, 1-10. https://doi.org/10.5815/ijmecs.2020.01.01.

30. Z. Hu, I. A. Tereykovski, L. O. Tereykovska, V. V. Pogorelov: International Journal of Intelligent Systems and Applications, 9(10), 2017, 57-62. https://doi.org/10.5815/ijisa.2017.10.07.

31. Z. Hu, Y.V. Bodyanskiy, N.Ye. Kulishova, O.K. Tyshchenko: International Journal of Intelligent Systems and Applications, 9(9), 2017, 29-36. https://doi.org/10.5815/ijisa.2017.09.04.

32. Z. Hu, M. Ivashchenko, L. Lyushenko, D. Klyushnyk: International Journal of Modern Education and Computer Science, 13(3), 2021, 13-22. https://doi.org/10.5815/ijmecs.2021.03.02.

33. N. Shakhovska, V. Yakovyna, V. Chopyak: Mathematical biosciences and engineering, 19(6), 2022, 6102-6123 https://doi.org/10.3934/mbe.2022285.

34. I. Izonin, R. Tkachenko, P. Vitynskyi, K. Zub, P. Tkachenko and I. Dronyuk,: International Conference on Decision Aid Sciences and Application (DASA), 2020, pp. 326-330, https://doi.org/10.1109/DASA51403.2020.9317124.