


ORIGINAL

Artificial neural networks with better analysis reliability in data mining

Redes neuronales artificiales con mayor fiabilidad de análisis en minería de datos

Bahar Asgarova¹, Elvin Jafarov¹, Nicat Babayev¹, Allahshukur Ahmadzada¹ , Vugar Abdullayev¹, Khushwant Singh²

¹Azerbaijan State Oil and Industry University. Baku, Azerbaijan.

²University Institute of Engineering & Technology, Maharshi Dayanand University. Rohtak-124001, India, MDU, Rohtak-124001.

Cite as: Asgarova B, Jafarov E, Babayev N, Abdullayev V, Singh K. Artificial neural networks with better analysis reliability in data mining. LatIA. 2024; 2:111. <https://doi.org/10.62486/latia2024111>

Submitted: 21-02-2024

Revised: 07-05-2024

Accepted: 24-08-2024

Published: 25-08-2024

Editor: Prof. Dr. Javier González Argote 

ABSTRACT

If there are relatively few cases, semi-supervised learning approaches make advantage of a large amount of unlabeled data to assist develop a better classifier. To expand the labeled training set and update the classifier, a fundamental method is to select and label the unlabeled instances for which the current classifier has higher classification confidence. This approach is primarily used in two distinct semi-supervised learning paradigms: co-training and self-training. However, compared to self-labeled examples that would be tagged by a classifier, the real labeled instances will be more trustworthy. Incorrect label assignment to unlabeled occurrences might potentially compromise the classifier's accuracy in classification. This research presents a novel instance selection method based on actual labeled data. This will take into account the classifier's current performance on unlabeled data in addition to its performance on actual labeled data alone. This uses the accuracy changes in the newly trained classifier over the original labeled data as a criterion in each iteration to determine whether or not the selected most confident unlabeled examples would be accepted by a subsequent iteration. Naïve Bayes (NB) will be used as the basic classifier in the co-training and self-training studies. The findings indicate that the accuracy and categorization of self-training and co-training will be greatly enhanced by SIS. As compared to semi-supervised classification methods, it will enhance accuracy, precision, recall, and F1 score, according to the findings.

Keywords: Supervised Instance Selection (SIS); Data Mining; Meta-Learning; Algorithm Selection.

RESUMEN

Si hay relativamente pocos casos, el aprendizaje semisupervisado de datos no etiquetados para ayudar a desarrollar un mejor clasificador. Para ampliar el conjunto de entrenamiento etiquetado y actualizar el clasificador, un método fundamental consiste en seleccionar y etiquetar las instancias no etiquetadas para las que el clasificador actual tiene una mayor confianza de clasificación. Este enfoque se utiliza principalmente en dos paradigmas distintos de aprendizaje semisupervisado: co-entrenamiento y auto-entrenamiento. Sin embargo, en comparación con los ejemplos auto-etiquetados que etiquetados por un clasificador, los casos reales etiquetados serán más fiables. La asignación incorrecta de etiquetas a sucesos no etiquetados podría comprometer potencialmente la precisión del clasificador en la clasificación. Esta investigación presenta un novedoso método de selección de instancias basado en datos etiquetados reales. Esto tendrá en cuenta el rendimiento actual del clasificador en datos no etiquetados, además de su rendimiento en datos etiquetados reales. Esto utiliza los cambios de precisión en el clasificador recién entrenado sobre los datos etiquetados originales como criterio en cada iteración para determinar si los ejemplos no etiquetados más seguros seleccionados serán aceptados en una iteración posterior. Se utilizará Naïve Bayes (NB) como clasificador básico en los estudios de co-entrenamiento y auto-entrenamiento. Los resultados indican que la precisión y la categorización del auto-entrenamiento y el co-entrenamiento mejorarán mucho con SIS. En comparación con los métodos de clasificación semisupervisada mejorará la exactitud, la precisión, el recuerdo y la puntuación F1, según los resultados.

Palabras clave: Selección Supervisada de Instancias (SIS); Minería de Datos; Meta-Aprendizaje; Selección de Algoritmos.

INTRODUCTION

Techniques for reducing data have been widely used to reduce the preparation time and capacity requirements of the classifiers. Include/characteristic selection, data pressure, data 3D square accumulation, and instance selection make up data reduction. Data 3D shape and highlight selection contribute to reducing the number of dimensions in the data to be mined, which in turn reduces the amount of time and capacity required for preparation. Data pressure also contributes to a decrease in demands for stockpiling. For instance-based or languid classifiers, it is typical to use instance selection methods that focus on selecting delegate instances.

These classifiers save all of the preparation examples instead of creating classification models. When a discrete instance has to be grouped, the hidden instance is compared to discrete instances and assigned to a class of instances that is closest to the hidden instance. Since the classifier requires the collection of all preparation cases, each hidden case is compared, and the quantity of preparation instances determines the instance-based classifiers' capacity requirements as well as the amount of time needed to arrange subtle instances. For instance-based classifiers, a few instance-selection methods have been developed and are often used.

An Artificial Neural Network (ANN) mimics the human mind in its most basic form. A unique cerebrum is capable of adapting to novel situations and changing circumstances. The human mind is very adept at sifting through fuzzy, incomplete material and drawing its own conclusions from it. One may examine the handwriting of others, for example, but the way they write may not be exactly the same as how it is written. A child can distinguish between the orange in a circle and the condition of a ball. In fact, a baby as young as a few days old may recognize its mother by touch, voice, and scent.

This is capable of identifying a genuine person from a blurry picture. The complex organ that governs the whole body is the cerebrum. Even the most primitive creature's thinking is more capable than the finest PC. Not only can it manage the physical parts of the body, but it can also do more complicated tasks like learning, thinking, and other tasks that cannot be expressed physically. The most extraordinary supercomputers still fall short of creating an artificial thinking machine.

One of the most common data mining tasks is classification. It's a controlled learning process. Early research in the topic focused on developing a few special techniques, such as neural networks, instance-based classifiers, and decision trees, to create classification models. Metrics such as classification accuracy, preparation time, stockpile need, and model conceivability are some of the measures used to compare the performance of various processes.

High classification accuracy, improved capacity for speculation, and adaptability are provided by neural network (NN) classifiers; nonetheless, they need a significant amount of preparation time, additional space, intelligibility, and a consistent learning capacity. In the last ten years, data mining research disciplines have focused mostly on data reduction systems. Advances in data reduction techniques led to the development of a few data reduction algorithms that are capable of performing instance and characteristic selection. Typically, trait selection techniques are used to reduce the number of dimensions in the data, which helps to shorten the preparation time and space complexity of the classifiers.

Literature survey

P. S. Raju et al.⁽¹⁾ In the banking and retail industries, data mining and CRM are necessities. This article includes several tasks and applications of data mining that are useful for these kinds of enterprises. The banking and retail sectors are aware that data mining is an important process for fundamental leadership and provides points of interest when carried out in a targeted manner.

This research, written by Vidhate D. R.⁽²⁾, focuses on the different perspectives of experts about the development of mega bazaars and consumer response. At this, customer purchase behavior at the super bazaar is broken down using learning mining.

This work by Lalithdevi B et al.⁽³⁾ provides clarification on data mining on web log data. This provides a detailed explanation of the web use mining activities and technological developments used in each project. Examples of age approaches and data preparation are described. That is useful for coming up with route plans. Lastly, it is assumed that although some mining algorithms use specifically designed strategies, the majority use the successive example age strategy.

R. B. Bhaise⁽⁴⁾ In order to improve the training under consideration, the main focus of this work is education data mining. Using the sample data, the developer used K-Means or clustering algorithms.

This process is used to organize and decompose data from different measures. Based on the understudy's presentation during the test, they formed groups. The information generated when a mining technique is put into practice is very beneficial for both instructors and students. Rjeswari K et al.⁽⁵⁾ Association rule mining is useful for evaluating the performance of the understudies in the research. In this research, Weka equipment is used for data exploration. This paper's main goal is to predict how an understudy would do on the college exam based on characteristics such as task, participation, and the inside test. This study concluded that by making more efforts in their unit exam, participation, task, and graduation, the poor understudies' college results would be enhanced. This research provides an overview of several data mining strategies for therapeutic individuals in basic leadership Chaurasia V et al.⁽⁶⁾ This allows the experts to predict the proximity of heart disease. This essay made use of The J48 Decision Tree, Naive Bayes, Strategies for bagging in the detection of cardiac illness. Because it provides comprehensible categorization criteria, the stowing calculation is thus superior than others. In order to determine the instability of a stock cost list with two distinct vision focuses difference and heading HyupRoh⁽⁷⁾ presents half-and-half models with neural networks and time arrangement models. The results showed that ANN-time arrangement models can increase the predictive power for the bearing exactness and deviation points of view. These investigational results showed that there is room for improvement in the crossover model that was established for measuring the volatilities of stock value record time arrangement.

Kim⁽⁸⁾ presented a genetic computation-based instance selection method for financial forecasting using artificial neural networks (ANN). In addition to reducing the dimensionality of the data, transformative instance selection may also remove superfluous and dangerous instances. The ideal association loads between layers in an ANN are also found using the transformational pursuit technique. Even while the majority of newly designed learning frameworks assume that preparation sets are appropriately updated, this isn't really the case. There are undoubtedly many domains in which a large number of precedents speak to one class, while just a few speak to the other. In the unlikely event if 99 % of the data come from a single class, a learning computation for the majority of plausible problems won't be able to demonstrate progress above the 99 % exactness attained by the minor classifier that labels all of the data with the bigger part class.

Cano⁽⁹⁾ presented an example selection approach based on developmental computation, a flexible method derived from ongoing development and very useful for research and development. Using a variety of experimental analyses, it seems

Better instance reduction and improved classification accuracy may be achieved efficiently using developing calculation. Furthermore, the authors emphasized that a transformational calculation would be a noticeable and persuasive tool in the instance selection strategy. This is very trustworthy given the conclusion that the instance selection technique introduced in this paper—hereditary calculation based instance selection—is a potent and popular one.

Japkowicz⁽¹⁰⁾ discussed the effects of dataset irregularities. The author evaluated two resampling methods. Irregular resampling, also known as oversampling, included randomly resampling the tiny class until it had the same number of tests as the larger portion class. Under-examining the greater part class tests until their numbers matched the number of minority class tests was another strategy taken into consideration: irregular under-testing. She saw that both of the examination strategies worked, and she also saw that using the improved testing methods did not provide any glaring margin of error in the region under consideration. A few more applications have successfully made use of neural networks. They can speculate more, have more clamor power, and be more precise, but they are not a good fit for data mining. Some prominent commentators on the use of neural networks in data mining classification tasks point out that they need more preparation time, are not as comprehensible, and do not have a consistent learning capability. A few studies have been reported in the article to address the trendy topic.

Data mining using supervised instance selection (sis) for better classification

The below figure 1 shows the flow chart of supervise instance selection algorithm for better classification is done. In this initially, input data is trained. Next data is preprocessed using data pre processing block and optional features are applied to the preprocessed data. For the featured data network training is applied. Along with that testing is performed for the trained data.⁽¹¹⁾

After testing data is validated using validation trained neural network. Now, instance selection using SIS process is performed for the validate data. After this SIS process again data is trained, tested and validated. Compact trained neural network will saved the validated data and train regarding to neural network concept. At last accuracy is improved this trained neural network.^(12,13)

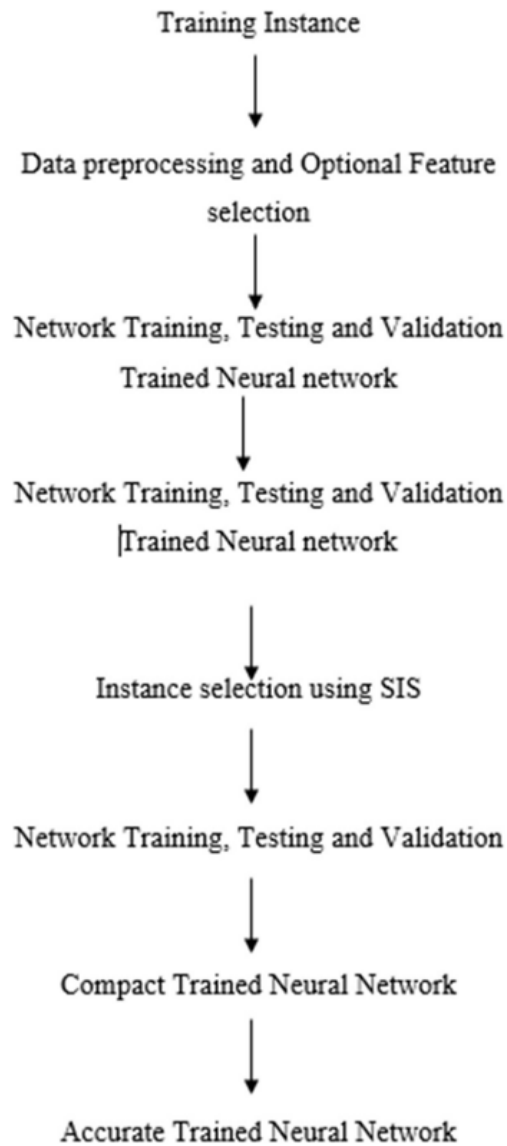


Figure 1. Flow Chart of supervised Instance Selection (SIS) for better classification

Algorithm

Step 1: in this initially, input data is trained.

Step 2: next data is preprocessed using data pre processing block and optional features are applied to the preprocessed data.

Step 3: for the featured data network training is applied. Along with that testing is performed for the trained data. After testing data is validated using validation trained neural network.

Step 4: now, instance selection using SIS process is performed for the validate data.

Step 5: after this SIS process again data is trained, tested and validated.

Step 6: compact trained neural network will saved the validated data and train regarding to neural network concept.

Step 7: at last accuracy is improved this trained neural network.

RESULTS

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Accuracy: it is defined as the ratio of correctly classified instances to the total predictions predicted by classifier.

Precision: the precision is defined as the ratio of number of instances which are classified as True Positives (TP) to the number of instances that are classified as False positive (FP+TP).^(14,15)

$$Precision = \frac{TP}{(TP + FP)}$$

Recall: it is also known as sensitivity, TPR (True positive Rate). It is a measure that gives the ratio of true positives. It is defined as the ratio of number of instances which are classified as TPs to the instances that are actually positive (i.e., TP + FP).^(16,17)

$$Recall = \frac{TP}{(TP + FN)}$$

F1-Score: this is a weighted average of precision and recall. For a better classification performance, the must be 1 and for bad performance, it must be zero.^(18,19)

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

The below table 1 shows the comparison table of Semi-supervised Classification Method and Supervised instance selection (SIS).^(20,21) In this accuracy, precision, recall and F1 score parameters are used. Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy, precision, recall and F1 score in effective way.^(22,23)

Table 1. Comparison table			
S.No	Parameters	Semi- supervised Classification Method	Supervised instance selection (SIS)
1	Accuracy	62 %	96 %
2	Precision	45 %	55 %
3	Recall	56 %	91 %
4	F1-Score	46 %	54 %

The below figure 2 shows the comparison of accuracy, recall for Semi-supervised Classification Method and Supervised instance selection (SIS). Compared with Semi- supervised Classification Method, Supervised instance selection (SIS) improves the accuracy and recall in effective way.^(24,25)



Figure 2. Comparison of Accuracy and Recall

The below figure 3 shows the comparison of precision for Semi-supervised Classification Method and Supervised instance selection (SIS). Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy and recall in effective way.^(26,27)

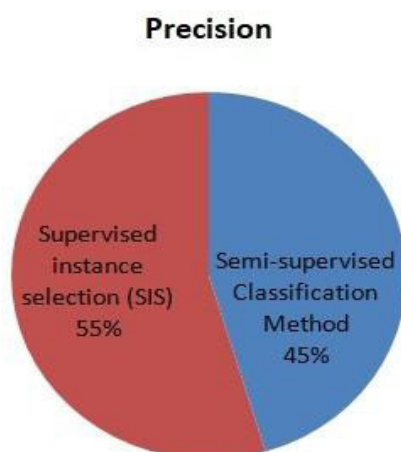


Figure 3. Comparison of Precision

The below figure 4 shows the comparison of F1-Score for Semi-supervised Classification Method and Supervised instance selection (SIS).^(28,29) Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy and recall in effective way.^(30,31)

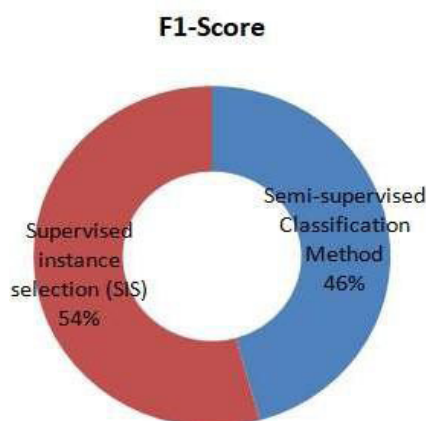


Figure 4. Comparison Of F1-Score

CONCLUSIONS

Because of the massive amounts of data that are continuously generated in a variety of study domains, the instance selection has increased recently. Thus, Supervised Instance Selection (SIS) data mining for improved artificial neural network classification accuracy yields useful results in this study. Based on the initial labeled data, a novel instance selection method is described here. As compared to semi-supervised classification methods, it will enhance accuracy, precision, recall, and F1 score, according to the findings.

BIBLIOGRAPHIC REFERENCES

1. Raju, P.S., Bai, V.R. &Chaitanya, G.K., 2014. Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1), pp.2650-2657.
2. Vidhate D. R. (2014), "A conceptual study of Consumer Behavior Analysis in Super Bazar using Knowledge Mining", *Sinhgad Institute of Management and Computer Application*, Pages : 70-75, ISBN : 978-81-927230-0-6.
3. Lalithdevi B., Ida A. M., Breen W. A. (2013), "A New Approach for improving World Wide Web Techniques in Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 1, Pages : 243-251, ISSN : 2277 128X.
4. Bhaise R. B. "An algorithm for a selective nearest neighbor decision rule", *IEEE Transactions on Information Theory*, Vol. 21, No. 6, pp.665-669.

5. Borkar S., Rjeswari K. (2013), “Predicting Students Academic Performance Using Education Data Mining”, International Journal of Computer Science and Mobile Computing, Volume 2, Issue 7, Pages: 273-279, ISSN: 2320-088X.
6. Chaurasia V., Pal S. (2013), “Data Mining Approach to Detect Heart Diseases”, International Journal of Advanced Computer Science and Information Technology, Volume 2, Issue 4, Pages : 56-66, ISSN : 2296-1739.
7. HyupRoh, “A compact and accurate model for classification”, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 6, pp.203-242 2007.
8. Kim, “Using neural networks for data mining”, Future Generation Computer Systems, Vol. 13, Nos. 2-3, pp.211-229, 2006.
9. Cano, “A general neural framework for classification rule mining”, Int. J. Computers, Systems and Signals, 2003 Vol. 1, No. 2, pp.154-168.
10. Japkowicz, “Symbolic interpretation of artificial neural networks”, 2000 IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 3, pp.448-463.
11. S. García, J. Derrac, J. R. Cano and F. Herrera, “Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 417-435.
12. Craven, M. and Shalvik, J. (1997) ‘Using neural networks for data mining’, Future Generation Computer Systems, Vol. 13, Nos. 2-3, pp.211-229.
13. Bhatia, S., Goel, A. K., Naib, B. B., Singh, K., Yadav, M., & Saini, A. (2023, July). Diabetes Prediction using Machine Learning. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1-6). IEEE. doi: 10.1109/WCONF58270.2023.10235187
14. Singh, K., Singh, Y., Barak, D., Yadav, M., & Özen, E. (2023). Parametric evaluation techniques for reliability of Internet of Things (IoT). International Journal of Computational Methods and Experimental Measurements, 11(2). <http://doi.org/10.18280/ijcmem.110207>
15. Singh, K., Singh, Y., Barak, D., & Yadav, M. (2023). Evaluation of Designing Techniques for Reliability of Internet of Things (IoT). International Journal of Engineering Trends and Technology, 71(8), 102-118. <https://doi.org/10.14445/22315381/IJETT-V71I8P209>
16. Singh, K., Singh, Y., Barak, D. and Yadav, M., 2023. Comparative Performance Analysis and Evaluation of Novel Techniques in Reliability for Internet of Things with RSM. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), pp.330-341. <https://www.ijisae.org/index.php/IJISAE/article/view/3123>
17. Singh, K., Yadav, M., Singh, Y., & Barak, D. (2023). Reliability Techniques in IoT Environments for the Healthcare Industry. In AI and IoT-Based Technologies for Precision Medicine (pp. 394-412). IGI Global. DOI: 10.4018/979-8-3693-0876-9.ch023
18. Singh, K., Singh, Y., Barak, D., & Yadav, M. (2023). Detection of Lung Cancers From CT Images Using a Deep CNN Architecture in Layers Through ML. In AI and IoT-Based Technologies for Precision Medicine (pp. 97-107). IGI Global. DOI: 10.4018/979-8-3693-0876-9.ch006
19. Kumar, S., Kumar, A., Parashar, N., Moolchandani, J., Saini, A., Kumar, R., Yadav, M., Singh, K., & Mena, Y. (2024). An Optimal Filter Selection on Grey Scale Image for De-Noising by using Fuzzy Technique. International Journal of Intelligent Systems and Applications in Engineering, 12(20s), 322-330. Retrieved from <https://ijisae.org/index.php/IJISAE/article/view/5143>
20. Singh, K., Singh, Y., Khang, A., Barak, D., & Yadav, M. (2024). Internet of Things (IoT)-Based Technologies for Reliability Evaluation with Artificial Intelligence (AI). AI and IoT Technology and Applications for Smart Healthcare Systems, 387. <http://dx.doi.org/10.1201/9781032686745-23>
21. Bhatia, S., Goel, N., Ahlawat, V., Naib, B. B., & Singh, K. (2023). A Comprehensive Review of IoT Reliability and Its Measures: Perspective Analysis. Handbook of Research on Machine Learning-Enabled IoT for Smart Applications Across Industries, 365-384. DOI: 10.4018/978-1-6684-8785-3.ch019

22. Singh, K., Mistrean, L., Singh, Y., Barak, D., & Parashar, A. (2023). Fraud detection in financial transactions using IOT and big data analytics. In *Competitivitatea și inovarea în economia cunoașterii* (pp. 490-494). <https://doi.org/10.53486/cike2023.52>
23. Sood, K., Dev, M., Singh, K., Singh, Y., & Barak, D. (2022). Identification of Asymmetric DDoS Attacks at Layer 7 with Idle Hyperlink. *ECS Transactions*, 107(1), 2171. <http://dx.doi.org/10.1149/10701.2171ecst>
24. Singh, K., Yadav, M., Singh, Y., Barak, D., Saini, A., & Moreira, F. Reliability on the Internet of Things with Designing Approach for Exploratory Analysis. *Frontiers in Computer Science*, 6, 1382347. doi: 10.3389/fcomp.2024.1382347
25. Singh, K., Yadav, M., Singh, Y., & Barak, D. (2024). Finding Security Gaps and Vulnerabilities in IoT Devices. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 379-395). IGI Global. DOI: 10.4018/979-8-3693-6016-3.ch023
26. Hajimahmud, V. A., Singh, Y., & Yadav, M. (2024). Using a Smart Trash Can Sensor for Trash Disposal. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 311-319). IGI Global. DOI: 10.4018/979-8-3693-6016-3.ch020
27. Yadav, M., Hajimahmud, V. A., Singh, K., & Singh, Y. (2024). Convert Waste Into Energy Using a Low Capacity Igniter. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 301-310). IGI Global. DOI: 10.4018/979-8-3693-6016-3.ch019
28. Singh, K., Yadav, M., & Yadav, R. K. (2024). IoT-Based Automated Dust Bins and Improved Waste Optimization Techniques for Smart City. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 167-194). IGI Global. DOI: 10.4018/979-8-3693-6016-3.ch012
29. Khang, A., Singh, K., Yadav, M., & Yadav, R. K. (2024). Minimizing the Waste Management Effort by Using Machine Learning Applications. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 42-59). IGI Global. DOI: 10.4018/979-8-3693-6016-3.ch004
30. Sharma, H., Singh, K., Ahmed, E., Patni, J., Singh, Y., & Ahlawat, P. (2021). IoT based automatic electric appliances controlling device based on visitor counter. DOI: [https://doi.org/10.13140/RG.2\(30825.83043\)](https://doi.org/10.13140/RG.2(30825.83043)).
31. Singh, K., & Barak, D. (2024). Healthcare Performance in Predicting Type 2 Diabetes Using Machine Learning Algorithms. In *Driving Smart Medical Diagnosis Through AI-Powered Technologies and Applications* (pp. 130-141). IGI Global. DOI: 10.4018/979-8-3693-3679-3.ch008

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Bahar Asgarova, Elvin Jafarov, Nicat Babayev, Vugar Abdullayev, Khushwant Singh.

Investigation: Bahar Asgarova, Elvin Jafarov, Nicat Babayev, Vugar Abdullayev, Khushwant Singh.

Methodology: Bahar Asgarova, Elvin Jafarov, Nicat Babayev, Vugar Abdullayev, Khushwant Singh.

Drafting - original draft: Bahar Asgarova, Elvin Jafarov, Nicat Babayev, Vugar Abdullayev, Khushwant Singh.

Writing - proofreading and editing: Bahar Asgarova, Elvin Jafarov, Nicat Babayev, Vugar Abdullayev, Khushwant Singh.