UDC 550.8.08
DOI 10.52171/herald.254

Artificial Intelligence-Based Lung Cancer Data Classification

N.A. Abilova[1], Y. Atay[2]

[1] *Gazi University, Graduate School of Natural and Applied Sciences (Ankara, Turkey)*
[2] *Gazi University, Computer Engineering (Ankara, Turkey)*

For correspondence:
Abilova Nazrin / e-mail: nazrin.ablv@gmail.com

Abstract

The article presents the results of an artificial intelligence-based study on the effectiveness of ensemble learning methods to improve accuracy in a lung cancer dataset. The results demonstrated that the Gradient Boosting, AdaBoost, LGBM, and SGD algorithms achieved the highest performance with an accuracy rate of 95.6%, while also providing strong precision, sensitivity, and F1-scores. Random Forest and XGBoost, with an accuracy of 91.3%, achieved successful results, proving their capacity to correctly distinguish between both classes. Overall, the ensemble methods used in this study exhibited strong performance in terms of both accuracy and generalization.

Keywords: ensemble learning, feature selection, classification, machine learning, artificial intelligence.

Submitted *29 April 2025*
Published *16 May 2025*

For citation:
N.A. Abilova, Y. Atay
[Artificial Intelligence-Based Lung Cancer Data Classification]
Herald of the Azerbaijan Engineering Academy, 2025 (online)

Süni intellektə əsaslanan ağciyər xərçəngi məlumatlarının sinifləndirilməsi

N.Ə. Əbilova[1], Y. Atay[2]

[1] *Qazi Universiteti, Təbiət Elmləri İnstitutu (Ankara, Türkiyə)*

[2] *Qazi Universiteti, Kompüter Mühəndisliyi (Ankara, Türkiyə)*

Xülasə

Məqalədə ağciyər xərçəngi məlumatlar toplusunda dəqiqliyi artırmaq üçün topluluq öyrənmə metodlarının effektivliyinə dair süni intellektə əsaslanan tədqiqatın nəticələri təqdim olunur. Nəticələr, Gradient Boosting, AdaBoost, LGBM və SGD algoritmalarının 95,6% doğruluqla ən yüksək performansı göstərdiyini və eyni zamanda yüksək precision, recall və F1-skorları təmin etdiyini göstərmişdir. Random Forest və XGBoost isə 91,3% doğruluqla uğurlu nəticələr əldə edərək, hər iki sınıf arasında düzgün ayrım etmə qabiliyyətini sübut etmişdir. Ümumilikdə, bu tədqiqatda istifadə olunan topluluq metodları həm doğruluq, həm də ümumiləşdirmə baxımından güclü performans nümayiş etdirmişdir.

Açar sözlər: topluluq öyrənməsi, xüsusiyyət seçimi, sinifləndirmə, maşın öyrənməsi, süni intellekt.

Классификация данных о раке легких на основе искусственного интеллекта

Н.А. Абилова[1], Й. Атай[2]

[1] *Университет Гази, Институт естественных наук (Анкара, Турция)*

[2] *Университет Гази, Компьютерная инженерия (Анкара, Турция)*

Аннотация

В статье приведены результаты исследования, основанного на искусственном интеллекте, по изучению эффективности методов ансамблевого обучения для повышения точности в наборе данных по раку лёгких. Результаты показали, что алгоритмы Gradient Boosting, AdaBoost, LGBM и SGD продемонстрировали наивысшую производительность с точностью 95,6% и также обеспечили высокую точность, чувствительность и F1-оценки. Random Forest и XGBoost достигли успешных результатов с точностью 91,3% и продемонстрировали способность правильно различать между двумя классами. В целом, методы ансамблевого обучения, использованные в этом исследовании, продемонстрировали высокую производительность как в точности, так и в обобщении.

Ключевые слова: ансамблевое обучение, отбор признаков, классификация, машинное обучение, искусственный интеллект.

Introduction

One of the greatest challenges in machine learning and data mining is feature selection. Feature selection is an important step in improving model accuracy, as each feature can affect the model's output and potentially lead to overfitting. Particularly when working with high-dimensional data, selecting unnecessary features can increase processing time and reduce model accuracy. In this context, performing feature selection helps identify the most important information, making the data more meaningful. Furthermore, the combination of different feature selection techniques through ensemble learning methods allows for stronger and more reliable results. Ensemble learning methods combine various feature selection techniques, enabling the attainment of more robust and dependable outcomes [1].

This approach can be likened to a scenario where multiple doctors evaluate a patient by considering different opinions and areas of expertise. Instead of relying on the decision of a single doctor, utilizing each one's distinct perspectives and experience helps in creating a more accurate and reliable treatment plan. Similarly, ensemble learning combines the weaknesses of each model to produce stronger and more accurate results [2].

This study examines how ensemble methods can enhance accuracy and yield more reliable results in classification tasks, particularly in biomedical datasets such as lung cancer data. In this context, each of the ensemble methods used contributes significantly to improving the overall success of the model and achieving more accurate classifications. Through the analyses conducted, the ability of each method and model to better distinguish between the features and classes within the dataset will be tested, and the approach that provides the best performance will be identified.

Analysis of literary sources

In recent years, ensemble learning methods have been considered an effective technique, especially for dealing with data imbalances. For example, Shah et al. tested an ensemble technique using the XGBoost method to predict traffic accidents and reported that the results accurately predicted the significant factors affecting traffic accidents [3]. Amgad et al., in their study on breast cancer, used a combination of ensemble and deep learning methods to enhance the performance of CNN-based models, showing that the ensemble approach performed better than individual models [4]. Enriquez et al. explored various fusion approaches in natural language processing, comparing the performance of voting, Bayesian combination, bagging, stacking, feature subspace creation, and cascading methods. In their experiments, stacking and cascading methods achieved good accuracy rates in all cases [5]. Opitz and Maclin presented bagging and boosting methods as ensemble techniques for neural networks and decision trees. Their research concluded that boosting methods outperformed bagging methods in terms of performance for a single classifier [6]. These studies demonstrate how ensemble learning methods are effectively applied across various fields and the advantages they offer over traditional methods. It is widely accepted that ensemble techniques, by combining different classifiers, provide more accurate predictions and reduce overall errors.

Setting the problem

In this study, two different datasets obtained from the CuMiDa (Curated Microarray Database) were used. CuMiDa provides 78 manually selected cancer microarray datasets, which were compiled from approximately 30,000 studies in the Gene Expression Omnibus (GEO) database **{** CITATION AnE25 \l 1068 **}**. The first dataset used in the study, Lung GSE19804, consists of 114 samples and 54,676 genes, divided into normal and tumor classes. The second dataset, Lung GSE18842, includes 90 samples and 54,676 genes, and it is also divided into normal and tumor (cancerous tissue) classes.

In this study, an ensemble method is proposed to improve accuracy, particularly for lung cancer data. The ensemble learning method is a technique that combines different types of models to achieve more accurate and reliable results. The ensemble methods used in the study combine the strengths of each model, ensuring accurate predictions.

Bagging (Bootstrap Aggregating) is the first ensemble method used in this study. Bagging divides a dataset into random subsets, and independent models are trained on each subset. The outputs of these models are then combined using majority voting or averaging. The Bagging method is particularly used to reduce overfitting, especially in models with high variance. It is especially beneficial for complex and variable models like decision trees.

Figure 1 illustrates the general working principle of the Bagging method. The figure clearly shows that the dataset is randomly split into subsets, with training performed on each subset, and the final prediction is made by majority voting. This approach helps balance the errors of each model by training them independently, thereby improving overall accuracy [1].

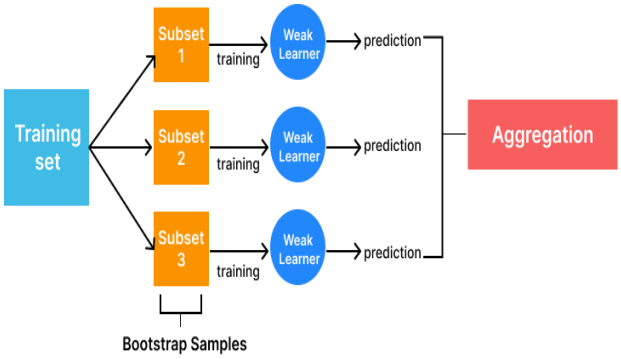

Figure 1 – Working Principle of the Bagging Method

Boosting is another important ensemble method in which weak learners are trained sequentially. In this method, each new model is designed to correct the errors of the previous model. Boosting primarily works to reduce bias errors and continuously improves performance by having each new model attempt to correct the errors of its predecessor, thereby increasing the accuracy of the model. However, boosting methods carry the risk of overfitting, so they must be used with caution. Boosting enables the creation of stronger models from weak learners, with each model targeting different errors during the process [1].

Figure 2 illustrates the general working principle of the Boosting method. The figure clearly shows how each new model sequentially corrects the errors of the previous model, resulting in the creation of a more powerful classifier.

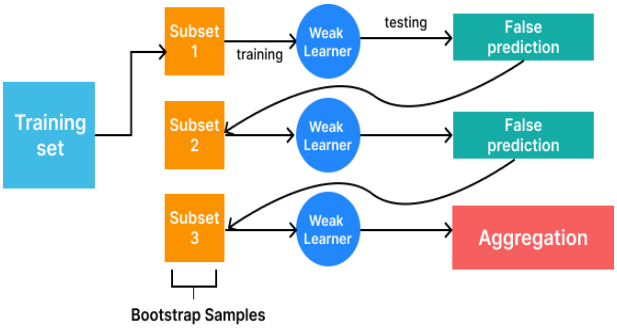

Figure 2 – Working Principle of the Boosting Method

The proposed method

In this study, among the proposed methods are Random Forest, Gradient Boosting, AdaBoost, LightGBM, SGD, and XGBoost. Each of these methods aims to improve classification performance by using different techniques and approaches.

The Random Forest algorithm uses the bagging technique to create multiple decision trees and combines the results of these trees using majority voting. Developed by Breiman, Random Forest increases the model's diversity by using random subsets and feature randomness during the training of each decision tree. This reduces the risk of overfitting. Only randomly selected features are used in the training of each tree, which reduces correlations and enhances the diversity of the trees. The overall accuracy of the model is obtained through the majority voting of the independent trees [8]. The Random Forest algorithm can also work effectively with missing data and typically performs well on high-dimensional datasets.

Gradient Boosting is a boosting algorithm that uses decision trees as base learners to create a powerful classifier. This algorithm adds a new model at each iteration to correct the errors of the previous model. Developed by Friedman, the goal of this algorithm is to make more accurate predictions by minimizing loss [9].Mathematically, each new model is added by focusing on the errors of the previous model as follows [10]:

$$Fm(x) = F_{m-1}(x) + \rho_m h_m(x) \qquad (1)$$

where F_{m-1} represents the predictions of the previous model, $h_m(x)$ represents the predictions of the new model, and ρ_m, determines the weight of the model.

AdaBoost (Adaptive Boosting) is a boosting algorithm developed by Freund and Schapire that creates strong classifiers using weak learners [11]. In this algorithm, at each iteration, new models are trained by giving more weight to misclassified examples. In this way, each new model attempts to correct the errors of the previous model. AdaBoost typically uses simple models, such as decision trees, to minimize the errors of each weak learner. The basic formula of AdaBoost is as follows [1]:

$$D_t(i) = D_{t-1}(i) \cdot \exp(-\alpha_t \cdot y_i \cdot h_t(x_i)) \qquad (2)$$

where $D_t(i)$ represents the weight of the i-th example, α_t represents the importance of the classifier at the t-th iteration, $h_t(x_i)$ represents the prediction of the model at the t-th iteration. Finally, a strong classifier is created by combining all the weak classifiers. However, AdaBoost is sensitive to noisy data and outliers and may carry the risk of overfitting.

XGBoost (Extreme Gradient Boosting) is a decision tree-based ensemble algorithm that creates strong classifiers using gradient boosting techniques. Developed by Chen and Guestrin [1], XGBoost prevents overfitting by adding a regularization term to the loss function. XGBoost improves learning by

correcting the model's errors at each iteration while using a second-order Taylor approximation, providing more precise and faster learning. This makes it suitable for fast and efficient application on large datasets.

LightGBM is an ensemble learning algorithm developed by Microsoft to overcome the efficiency and scalability challenges associated with high-dimensional data and large datasets, which are present in XGBoost. This method uses techniques such as exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS) to increase the accuracy of the model while processing data more quickly and efficiently. These features make LightGBM particularly effective and computationally efficient on large datasets [1].

Stochastic Gradient Descent (SGD) is a machine learning algorithm commonly used for fast learning on large datasets. This algorithm speeds up the learning process by updating the model's weights based on a randomly chosen example of the data at each iteration. Developed by Friedman, SGD can also produce effective results on nonlinear problems. SGD is successfully used in algorithms like support vector machines and aims to optimize linear loss functions [12].

Each algorithm has its advantages and challenges depending on the specific dataset, but the combination of ensemble methods provides a strong approach to increasing accuracy and preventing overfitting. Ensemble methods not only create stronger models but also enhance the model's ability to generalize, allowing for more robust and reliable results.

Experimental studies

In the experimental studies, the data from both the GSE18842 and GSE19804 datasets [7] were split into training and test sets. Each dataset was randomly divided into two groups using the train_test_split(X, y, test_size=0.20) function, with an 80% training and 20% test data ratio. This separation was done to test the models' generalization ability.

Experiments on the GSE19804 dataset. In the analysis of the Lung GSE19804 dataset, the data is divided into normal and tumor classes. Upon examining the data, it is observed that there is a significant difference between the normal and tumor classes, with the majority of the data belonging to the normal class [7].

In the analysis of the GSE19804 dataset, the performance of various classification algorithms on lung cancer data was evaluated. Random Forest achieved 0.91 accuracy, with precision, recall, and F1-score values of 0.94 for the normal class and 0.94 for the tumor class. The AUC on the ROC curve was 0.99, indicating strong performance. Gradient Boosting, AdaBoost, LGBM, and SGD all reached 0.96 accuracy, with precision, recall, and F1-score values ranging from 0.94 to 1.0 for both classes. The AUC values were 0.97, 0.99, 0.98, and 0.99, showing high overall performance. XGBoost also achieved 0.91 accuracy, with AUC values and class separation similar to Random Forest. The Learning Curve analysis revealed increasing accuracy over time, with improved generalization capacity as the number of training examples grew.

When the results presented in Table 1 are evaluated overall, the accuracy of the RandomForest model was calculated to be 0.91, with an F1-score of 0.9375, and precision and recall values of 0.9375, respectively. The GradientBoosting, AdaBoost, LGBM, and SGD models each demonstrated the best performance with 0.96 accuracy. In these models, the F1-score was 0.9697, and the precision and recall values ranged between

0.9412 and 1.0. These results show that these models performed excellently in predicting the tumor class and minimized some errors in the normal class. XGBoost, on the other hand, exhibited similar performance to RandomForest, with 0.91 accuracy, 0.9375 F1-score, and 0.9375 precision and recall values. Overall, GradientBoosting, AdaBoost, LGBM, and SGD stand out as models with the highest accuracy and strong performance, while RandomForest and XGBoost achieved lower accuracy with their results.

Table 1 – Performance evaluation of classification models on the GSE19804 dataset

Method	**Acc**	**F1**	**Pre**	**Rec**
RandomForest	0.91	0.94	0.94	0.94
GradientBoosting	**0.97**	**0.97**	**0.94**	**1.0**
AdaBoost	0.96	0.97	0.94	1.0
LGBM	0.96	0.97	0.94	1.0
SGD	0.96	0.97	0.94	1.0
XGB	0.91	0.94	0.94	0.94

Experiments on the GSE18842 dataset. The Lung GSE18842 dataset is divided into normal and tumor classes. The distribution between these classes is shown in the visual. Upon examining the data, it is observed that there is a significant difference between the normal and tumor classes, with the majority of the data belonging to the tumor class. This class imbalance could be an important factor to consider during the training of the model [7].

In the analysis of the GSE18842 dataset, the performance of different classification algorithms was examined. The Random Forest method achieved excellent results with 1.00 accuracy. The precision, recall, and f1-score values for both classes were calculated to be 1.00, and the AUC value on the ROC curve was 1.00. Throughout the training process, the model's accuracy increased and continued to produce correct results without overfitting. In the analysis using the Gradient Boosting method, the accuracy was calculated at 0.96, and it was observed that the model could distinguish both classes with high accuracy. The precision, recall, and f1-score values were set at 1.00, and the AUC value was also calculated to be 1.00, indicating that the model performed excellently for both classes. In the test using AdaBoost, the accuracy was calculated at 0.94, with precision, recall, and f1-score values of 0.86, 1.00, and 0.92 for the normal class, and 1.00, 0.92, and 0.96 for the tumor class, respectively. The AUC value on the ROC curve was 1.00, emphasizing that the model has a high capacity for correctly distinguishing both classes. In the LGBM method, the accuracy was calculated at 89%, with precision, recall, and f1-score values of 0.75, 1.00, and 0.86 for the normal class, and 1.00, 0.83, and 0.91 for the tumor class, respectively. The AUC value was 0.92, indicating that the model has a very good ability to distinguish between both classes but needs slight improvement to reach perfection.
In the analysis using the SGD method, the accuracy was calculated at 1.00, with precision, recall, and f1-score values of 1.00 for both classes. The AUC value on the ROC curve was 1.00, proving that the model could distinguish both classes excellently. Finally, in the XGBoost method, the accuracy was calculated at 1.00, with precision, recall, and f1-score values of 1.00 for both classes. The AUC value on the ROC curve was 1.00, indicating that the model was able to distinguish both classes with

high accuracy, and its overall performance was excellent.

When the results presented in Table 2 are evaluated overall, the accuracy of the RandomForest model was calculated to be 1.00, and the F1-score was also determined to be 1.0. Both the precision and recall values were 1.0, indicating that this model exhibited excellent performance in correctly predicting both classes. The accuracy of the GradientBoosting model was also calculated to be 1.00, with F1-score, precision, and recall values again being 1.0. This model also performed strongly by predicting both classes with high accuracy.

Table 2 – Performance evaluation of classification models on the GSE18842 dataset

Method	**Acc**	**F1**	**Pre**	**Rec**
RandomForest	**1.00**	**1.00**	**1.00**	**1.00**
GradientBoosting	**1.00**	**1.00**	**1.00**	**1.00**
AdaBoost	0.95	0.96	1.00	0.92
LGBM	0.89	0.91	0.94	0.83
SGD	**1.00**	**1.00**	**1.00**	**1.00**
XGB	**1.00**	**1.00**	**1.00**	**1.00**

AdaBoost achieved 0.94 accuracy, and the F1-score was calculated to be 0.96. The precision value was 1.0, while the recall value was determined to be 0.92. It can be said that this model performed well in predicting the tumor class, but there were some errors in the normal class. The accuracy of the LGBM model was calculated at 0.89, and the F1-score was 0.91. The precision value was 1.0, but the recall value was slightly lower at 0.83. This model exhibited some errors, particularly in the normal class, but overall, it showed good performance. The accuracy of the SGD model was calculated to be 1.00, and the F1-score, precision, and recall values were all 1.0. This model achieved excellent performance by correctly distinguishing both classes. Finally, XGB, like RandomForest, exhibited excellent performance with 1.00 accuracy, 1.0 F1-score, and 1.0 precision and recall values.

Comparison with methods in the literature. In the literature, various classification studies on lung cancer data have compared the success of different algorithms. These studies aim to obtain more robust and reliable results by combining the strengths of various model types. Some studies in the literature emphasize the importance of various factors affecting the performance of each model, such as class imbalance, high-dimensional datasets, and the selection of different features. The results obtained in our study, when compared to the findings in the literature, show that the models used in our study performed effectively on lung cancer data and that each model provided successful results across different metrics.

When compared to methods in the literature, the performance of the models used in this study achieved quite impressive results. Specifically, models such as RandomForest, Gradient Boosting, AdaBoost, LGBM, and SGD demonstrated strong performances with accuracy rates of 0.91 and 0.96, respectively, achieving high results in Precision, Recall, and F1-Score values. Particularly, Gradient Boosting, AdaBoost, LGBM, and SGD models each performed excellently with 0.96 accuracy, drawing attention with Precision values of 0.94 and 1.0 Recall values. These results indicate that high-accuracy predictions were made when compared to studies in the literature. Table 3 compares the performance of

classification models from the literature using the GSE19804 dataset. For example, FCBF, proposed by Abdelazim et al. [13], achieved an accuracy of 0.95, which is impressive but still falls short when compared to the performance of models used in this study. Similarly, the L1, LEN, and L1/2 methods proposed by Wu et al. [14] demonstrated accuracy rates ranging from 0.81 to 0.87, while the methods used in this study exhibited better performance. This shows that the ensemble methods employed in our study provided superior accuracy and reliability compared to existing methods in the literature.

The results obtained from the analysis of the GSE18842 dataset show that the models used provided very high accuracy and reliability. RandomForest, GradientBoosting, SGD, and XGB models demonstrated excellent performance with 1.00 accuracy, with precision, recall, and f1-score values of 1.0 for both classes, indicating that the model was able to perfectly distinguish both classes. The AdaBoost model achieved high performance with 0.94 accuracy, 0.96 f1-score, and 0.92 recall, marking significant success. Although LGBM had the lowest accuracy at 0.89%, it still performed quite well.

Table 3 – Performance comparison of literature classification models on the GSE19804 dataset

Method	**Acc**	**F1**	**Pre**	**Rec**
FCBF [13]	0.95	0.95	0.97	0.96
L1 [14]	0.81	-	0.87	-
LEN [14]	0.81	-	0.92	-
L1/2 [14]	0.87	-	0.92	-

When compared to methods in the literature, the performance of the models used in this study is quite remarkable. Table 4 shows the performance comparison of classification models from the literature using the GSE18842 dataset.

Table 4 – Performance comparison of literature classification models on the GSE18842 dataset

Method	**Acc**	**F1**	**Pre**	**Rec**
LAD [15]	0.98	-	0.82	-
PAM [16]	0.80	-	-	-

For example, the LAD method proposed by Bartosh & Masich **{**CITATION Bar22 \l 1068 **}** achieved 0.98 accuracy, while in our study, the RandomForest, GradientBoosting, SGD, and XGB models achieved higher success with 1.00 accuracy. Additionally, the PAM method proposed by Yu et al. [16] showed lower performance with 0.80 accuracy. The results indicate that the ensemble methods used in our study provided high accuracy and reliability, offering a very strong performance when compared to existing methods in the literature. These findings demonstrate that the methods used in our study were very successful on lung cancer data.

Overall, models such as GradientBoosting and SGD have produced stronger and more reliable results on datasets with class imbalance, such as lung cancer. In such datasets, it is observed that Boosting methods are effective in improving accuracy by correcting the low-performing examples. XGB, on the other hand, distinguished both classes accurately with high accuracy and stood out for its ability to learn quickly and provide high accuracy.

Results and discussion

In this study, the performance of various classification algorithms used on lung cancer data was evaluated. Specifically, models such as Random Forest, Gradient Boosting, SGD, and XGBoost achieved excellent results with 100% accuracy and correctly distinguished both classes. The precision, recall, and f1-score values of these models were also high, with particularly strong performance on the tumor class. On the other hand, AdaBoost and LGBM models achieved lower accuracy rates (0.94 and 0.89, respectively), but still demonstrated significant success. Some errors were observed in the normal class of the AdaBoost model, while the accuracy deficiency in the normal class of the LGBM model was noticeable.

Overall, the ensemble methods used provided high accuracy and reliability, with particularly effective performance on the tumor class. Random Forest, Gradient Boosting, SGD, and XGBoost models stood out in terms of both accuracy rate and class prediction success. Future work offers new opportunities to further improve these models and enhance their generalization capacities by using different datasets and hyperparameter optimizations. Specifically, testing the model's generalization capacity through experiments with larger datasets could help make machine learning applications in healthcare, such as lung cancer, more effective. Additionally, applying more advanced techniques to deal with noisy data and improve data imbalance is recommended.

Conflict of Interests

The authors declare there is no conflicts of interests related to the publication of this article.

REFERENCES

1. **Ibomoiye D.M., Yanxia S.** A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access. 2022. vol. 10. Pp. 99129–99149.
2. **Yu S., Xuewen L., Jing Z., Zhanli L.** Classifier selection and ensemble model for multi-class ımbalance learning in education grants prediction. Applied artificial ıntelligence. 2021. vol. 35, no. 4. Pp. 290-303.
3. **Milind S., Kinjal G., Kinjal A. P., Harsh K., Rohini P., Ankita K.** Theoretical evaluation of ensemble machine learning techniques. 25th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, 2023.
4. **Nadeen A., Mariam A., Haidy H., Moamen Z., Ammar M.** A robust ensemble deep learning approach for breast cancer diagnosis. Intelligent methods, systems, and applications (IMSA), Giza, 2023.
5. **Fernando E., Fermín L. C., F. Javier O., Carlos G. V., José A. T**. A comparative study of combination applied to NLP tasks. Information fusion, 2013. vol. 14, no. 3. Pp. 255-267.
6. **David O., Richard M.** Popular ensemble methods: An empirical study. JAIR. 1999. vol. 11, no. 1. Pp. 169-198.
7. **Breiman Leo.** Random forests. Machine learning. 2001. vol. 45, no. 1. Pp. 5-32.
8. **Breiman Leo.** Arcing classifiers. The annals of statistics. 1998. vol. 26, no. 3. Pp. 801-824.
9. **Wenyang W., Dongchu S.** The ımproved adaboost algorithms for ımbalanced data classification. Information sciences. 2021. vol. 563, no. 6. Pp. 358–374.

10. **Yoav F., Robert E.S.** A short Introduction to boosting. Journal of japanese society of artificial ıntelligence. 1999. vol. 14, no. 5. Pp. 771-780.
11. **Léon Bottou.** Large-scale machine learning with stochastic gradient descent. in proceedings of COMPSTAT'2010. 2010.
12. An extensively curated microarray Database. https://sbcb.inf.ufrgs.br/cumida. [Accessed 20 04 2025].
13. **Waleed M.E., Marwa A.A., Mona M.N.** Feedforward deep learning optimizer-based RNA-Seq women's cancers detection with a hybrid classification models for biomarker discovery. International Journal of Advanced Computer Science and Applications (IJACSA). 2022. vol. 13, no. 12.
14. **Shengbing W., Hongkun J., Haiwei S., Ziyi Y.** Gene selection in cancer classification using sparse logistic regression with L1/2 regularization. Applied Sciences. 2018. vol. 8, no. 9. Pp. 1569.
15. **Bartosh M., Masich I.** Cancer prediction models using gene expression and logical analysis of data. Hybrid Methods of Modeling and Optimization in Complex Systems (HMMOCS 2022), Krasnoyarsk, 2022.
16. **Hui Y., Qinghua X., Fang L., Xun Y., Jialei W., Xia M.** ıdentification and validation of long noncoding RNA biomarkers in human non-small-cell lung carcinomas. 2015, vol. 10, no. 4. Pp. 645-654.