Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

Medical decision support systems for diagnosing diseases based on ensemble learning algorithms

Luma Fawaz Jarallah

College of Administration and Economics Department of Management Information Systems, Mosul University, Iraq
Email: luma_fawaz@uomosul.edu.iq

Article information

Article history:
Received:28/5/2024
Accepted:1/10/2024
Available online:15/12/2024

Abstract

Diagnosing diseases in humans is the first step in treating diseases, and knowing it is important to determine treatment and deal with the disease in the correct way. Diagnosis is made in medical institutions using available tools and specialists in each medical field to determine the problem presented by the patient. Modeling and analysis of medical data is important in healthcare and social applications in areas related to disease prediction and diagnosis. The model selection strategy is an important determinant of the performance and acceptance of a medical diagnostic decision support system. This paper proposes a stacked learning model derived from multiple ensembles learning algorithms, including Random Forest, Catboost and XGBoost. To determine the effectiveness of the model, it was tested using eight data sets covering different diseases to help make disease diagnosis decisions. The results show that the proposed model generally outperforms individual machine learning models in terms of accuracy.

Keywords:
Medical decision support, disease diagnosis, machine learning, Ensemble learning.

Correspondence:
Author: Luma Fawaz Jarallah
Email: *luma_fawaz@uomosul.edu.iq*

I. Introduction

Medical data analysis and decision making are areas that are growing day by day. With the development of technology and the continuous accumulation of medical data, more and more medical facilities have begun to pay attention to data analysis and finding patterns [1-3]. By applying artificial intelligence, medical institutions can better understand the nature of diseases and medical needs and improve the clinical decision-making process [4-5]. The health sector faces many challenges, including resource management, disease prediction and diagnosis, medical quality control, and the patient's treatment process [3][6]. Artificial intelligence technologies can help health sector managers better understand patient needs and disease patterns and make accurate disease prediction and management [4]. Using trained models, it is possible to predict the disease risks of different groups of people based on previous patient data and disease development trends and formulate corresponding prevention and treatment strategies [7]. This not only improves patient outcomes but also reduces hospital medical costs [2][7]. Diagnosing and predicting diseases is an essential part of a doctor's work, and making the correct clinical decision is crucial to the patient's diagnosis and treatment results [8].

In recent years, researchers have conducted many studies to diagnose various diseases using machine learning and have reached promising results in this field [9]. Machine learning (ML) algorithms learn from data, which is a self-training process in terms of structure, as it performs fitting during the learning process. Machine learning has been successfully applied in almost every field from financial sector, education, meteorology to healthcare. In healthcare,

machine learning methods are mainly used for the purpose of diagnosing diseases. Studies show that using machine learning techniques to diagnose diseases is a promising solution. Diagnosing the disease early can help treat the patient and significantly improve the health condition [10-12]. The traditional way of making the diagnosis is usually expensive and time-consuming. Furthermore, studies have shown promise for using time- and money-efficient machine learning-based methods to diagnose diseases. Both common and uncommon diseases can now be diagnosed using machine learning methods. This paper aims to develop a model for diagnosing problems using the concept of stacking based on the aforementioned models such as Random Forest, Catboost, and XGBoost. These algorithms were chosen because of their efficiency in managing nonlinear correlations within data sets. By applying the proposed architecture to different disease-related datasets, we will attempt to demonstrate that this ensemble-based approach is not only a way to significantly increase overall diagnostic performance, but also offers a more accurate and useful tool to support the work of clinicians. Decisions from each machine learning model individually.

1.1 Bagging

Bagging is one of the ways to improve machine learning models by collecting more than one model [13]. It has been widely used in many aspects. Its principle can be described as a linear combination of multiple functions, which is suitable for optimizing some algorithms with relatively poor accuracy. Often a new learning algorithm can be obtained that significantly improves accuracy. In short, it is a weak model set and the training set determined after multiple trainings meets the specified conditions, the corresponding function can be A certain sequence of values is obtained, and the final evaluation function is formed through mathematical processing of the function values [14].

Weak learning algorithms refer to classification algorithms that are independent and do not require high accuracy. The bagging-first algorithm improves the generalization ability of each weak classifier after fusion. Generalization ability reflects the classifier's ability to adapt to new things. The stronger the ability to generalize, the stronger its ability to adapt to new things. Second, it improves the accuracy and other evaluation indicators of the model. It can be seen that using the bagging algorithm to solve the disease diagnosis problem is quite an effective method [13][15].

In this paper, the heterogeneous ensemble learning method is used in the bagging algorithm, which is formed by merging several weak classifiers of different types of algorithms. This is because the difference algorithm itself has certain differences. After merging, the generated classification decision boundaries will be different, that is, they will make different errors when making decisions. After merging, sharper boundaries can often be obtained, thus the overall classification errors can be reduced and better prediction results can be achieved.

1.2 Boosting

Boosting: a ladder-like modeling method. Models are trained in order, and the training set of the base model undergoes a certain change each time according to a certain strategy [15]. A linear combination of the predictive scores of the baseline forecast produces the predicted outcome. Boosting is a powerful learning method, as it combines several "weak" classifiers to produce a strong set of classifiers. The weak classifier performs slightly better than random selection, so it can be designed to be very simple and computationally inexpensive. Several weak classifiers are combined to form an integrated strong classifier [16].

1.3 Stacking

Stacking means that the final prediction result of the integrated model is jointly determined by a number of different trained models, so that the best result is achieved. The general idea is to combine the results of different trained models, ensuring diversity of results. This method aims to integrate the results of different algorithms, because the results of different algorithms have different concentrations and can meet multiple real results in classification [17-18].

In binary classification, the set of models is trained separately and then the trained models are used to generate predictions and create a new data set whose inputs are the predictions of the different models while its output is the true values that the model should predict. A meta-model is used and trained using the new data set to determine the most accurate models in predicting certain cases. This means that the characteristics of a specific model and the extent of the strengths, weaknesses and strengths of each model can be learned according to the case.

II. Related Studies

Many researchers have developed machine learning models to diagnose diseases. Muhammad et al. (2020) The authors used a set of machine learning algorithms to determine the best model: DT, NB, in addition to sequential optimization (SMO). The models were applied to two data sets: WBC and breast cancer. The researchers focused on the data imbalance and were balanced. The researchers concluded that their proposed algorithm achieved better results than DT and NB. Despite the good results, data balancing and data reshaping procedures may affect the performance of the models [19]. In the study of Ma et al. (2020) and others. The researchers proposed the use of the Heterogeneous Modified Artificial Neural Network (HMANN) in the detection and diagnosis of chronic kidney disease. The study found the superiority of the model compared to traditional methods, but it did not address the effectiveness of the model in diagnosing other diseases [20]. In Assegie (2021), researchers proposed using network search to determine the best nearest neighbor (KNN) settings in breast cancer detection. The results showed that modifying the number of neighbors parameter had a significant impact on

the model's accuracy. They demonstrated that by fine-tuning the settings, it is possible to obtain a 4% increase in accuracy compared to the default parameters of the model [21]. Using a population of 802 persons with SLE or other rheumatologic disorders, Adamichou et al. developed a diagnosis algorithm in 2021 based on random forest and least absolute shrinkage and selection sperator-logistic regression. The SLE Risk Probability Index scoring system, which has an accuracy of 94.2% for scores greater than 7, was developed using this algorithm, which demonstrated excellent performance in the diagnosis of SLE (accuracy of 94.8%) and, more importantly, high sensitivity in the early diagnosis of the disease (sensitivity of 93.8%) [22]. In comparison to previous heart disease datasets, the dataset from Ahmad G. N. et al.'s study was bigger because to the combination of datasets from Cleveland, Hungarian, Switzerland, Statlog, and Long Beach VA. They examined the effectiveness of the following algorithms for classifying cardiac diseases: LR, KNN, SVM, Nu-Support Vector Classifier (Nu-SVC), DT, RF, NB, ANN, AdaBoost, Gradient Boosting (GB), Linear Discriminants Analysis (LDA), and Quadratic Discriminant Analysis (QDA). The authors of this study said that the RF algorithm produced the greatest classification accuracy [23]. A random forest-based machine learning method was created in 2022 by Ma et al. [24] to distinguish SLE patients from healthy individuals. Based on changes in the subpopulation of peripheral blood mononuclear cells in SLE patients, this mathematical model was created. A LASSO-LR model-based method was created by Han et al. in 2023 [25] that could distinguish between SLE and SLE-SS overlap as well as forecast the likelihood of SLE-to-SLE-SS progression. To create the prediction model, the program determined the top five attributes: watery eyes, dry mouth, anti-Ro52, anti-SSB, and positive for rheumatoid factor.

Despite the superiority of the models proposed by researchers, the studies that used a model and tested it for more than one disease are rare. Therefore, the performance of the models cannot be generalized if more than one disease or data set is diagnosed.

III. DATA AND METHODS

3.1 Experimental data sets

The dataset used in this study is eight datasets covering a spectrum of diseases. The Wisconsin Breast Cancer dataset was used in the UCI Machine Learning Repository [26], and for heart disease analysis, the CARDIOVASCULAR dataset was used [27]. To predict risk factors associated with diabetes, the Pima Indian dataset, made available by the National Diabetes Institute at Johns Hopkins University, was used [28]. To predict kidney failure, the Chronic Kidney Disease data set was used [29]. The Indian Liver Disease Classification Database (ILPD; Indian Liver Patients Data Set) was used [data 5]. To diagnose multiple sclerosis, the Multiple Sclerosis Disease data set was used [30]. To diagnose Parkinson's disease, a database taken from the University of California, Irvine (UCI) machine learning repository was used [31]. Finally, to distinguish thyroid diseases, the Thyroid Disease dataset was used [32]. Table 1 describes the datasets in terms of number of rows and columns, data types, and features.

3.2 Data preparation

Data preparation is a basic process in data modeling and aims to prepare data for use in analysis processes and building predictive models. This study uses the Pandas software package to analyze data quality. Data preprocessing and modeling are implemented in Python 3.10.2. sklearn is used to create training sets and test sets. The Pandas library is used to read the CSV file containing the data. Reading the data follows by finding the missing values in the data using the average values for each column. This is done by the `fillna` function in Pandas. The formatted and processed data is returned for use in analysis and model building.

Table 1. Datasets Description.

Dataset	***Data Types***	***Description***
Breast [26]	float64, object	Contains data related to breast cancer, with 569 rows and 31 columns.
Cardio [27]	int64, float64	Contains data related to cardiovascular health, with 70,000 rows and 12 columns.
Diabetes [28]	int64, float64	Contains data related to diabetes with 768 rows and 9 columns.
kidney_disease [29]	int64, float64, object	Contains data related to kidney diseases, with 397 rows and 26 columns.
liver_patient [30]	int64, float64	Contains data related liver patients, with 579 rows and 11 columns.
multiple_scleros is [31]	int64, float64	Contains data related to multiple sclerosis, with 125 rows and 20 columns.
Parkinsons [32]	int64, float64	Contains data related to Parkinson's disease, with 195 rows and 23 columns.
thyroidDF [33]	int64, float64, object	Contains data related to thyroid conditions, with 20 rows and 30 columns.

3.3 Data modeling

In this study, a set of machine learning algorithms were used in addition to the proposed model. For comparison, models were used that varied between Bagging and Boosting, in addition to the proposed architecture.

Stacking is a model fusion algorithm, also called model stacking, the basic idea is to train several models first from the initial training set, and then combine the output results of one model as sample features, and label the original samples as new data samples. Then a new training set is created, a new model is trained based on the new training set, and finally the new model is used to predict the sample. The advantage is to reduce the generalization error of a single model. The steps to design this study are as follows:

1. First, the subset of selected features is divided into two parts, part including 20% sample data for testing and 80% for training.

2. The first layer stacking model can be understood as an initial prediction layer. In this layer, three powerful models were chosen, namely: Random Forest (RF), Catboost (CT) and XGBoost. These three tree models are all powerful models and are all based on decision trees.
3. The three models in the first layer were trained using the four-fold cross-validation method. The idea of four-fold cross-validation is to divide the data into 4 parts, train 3 parts of the data, each part of the data will be used as a validation set, and then average the 4 results obtained. After passing the first layer of training, the three prediction results of the three models are combined, which is used as input to the second layer model for training. The basic model for the second layer uses a decision tree classifier. The specific stacking model merging steps are shown in **Fig 1**.

IV. Experimental Results

The sample set collected in this paper contains a total of 8 disease types. The target result for each sample contains one disease. When evaluating and analyzing the target samples, in order to ensure the effectiveness of the bagging integration algorithm, a group of ensembles learning algorithms were adopted and compared with the proposed model. Finally, three indicators (accuracy, precision, recall) were used in formula (1), formula (2), and formula (3).

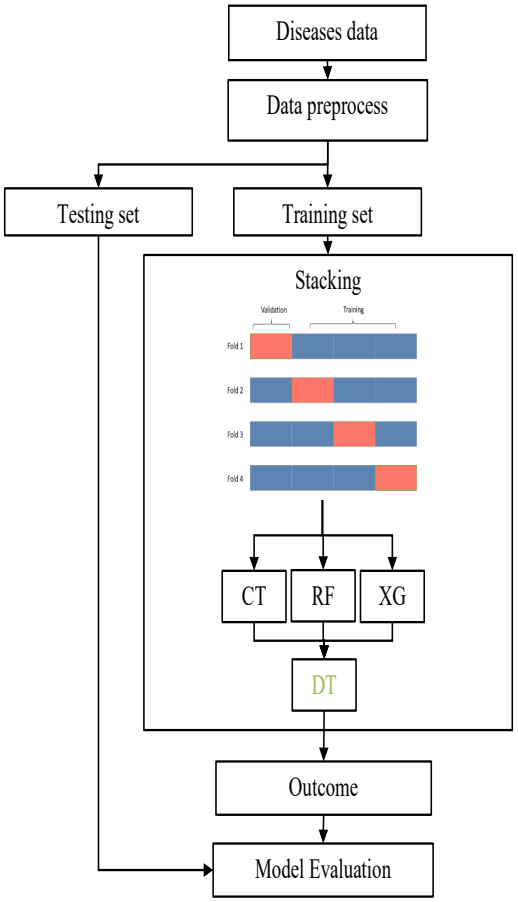

Fig 1. Flow chart of the proposed stacking model

$$\text{Accuracy} = TP + TN / (TP + FP + FN + TN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

Where:

TP (True Positive) denotes a truly positive response.
TN (True Negative) an actual negative response.
FP (False Positive) a false positive response.
FN (False Negative) a false negative response.

Fig 2, 3, 4 compares the ensemble learning model.

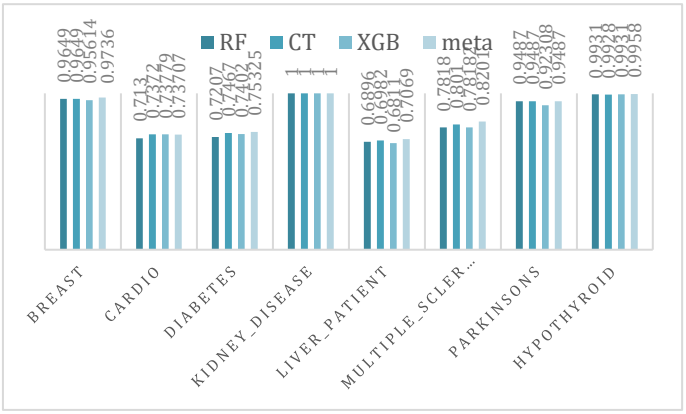

Fig. 2. Comparison based on accuracy

It is noted from **Fig 2** that the proposed meta- ensemble learning (MEL) model achieved the highest accuracy in the case of the data sets (Breast, Diabetes, Liver Patient, Hypothyroidism, and Multiple Sclerosis), which were (0.9736, 0.75325, 0.7069, 0.9958, and 0.8201), respectively. While in the Cardio dataset the accuracy was close between the models, but XGBoost had the highest value of 0.73779. At the same time, in Parkinson's Dataset, the highest accuracy is 0.9487 for both the Random Forest and the meta model. Finally, in the case of Kidney Disease Dataset, all models obtained an accuracy of 1.0.

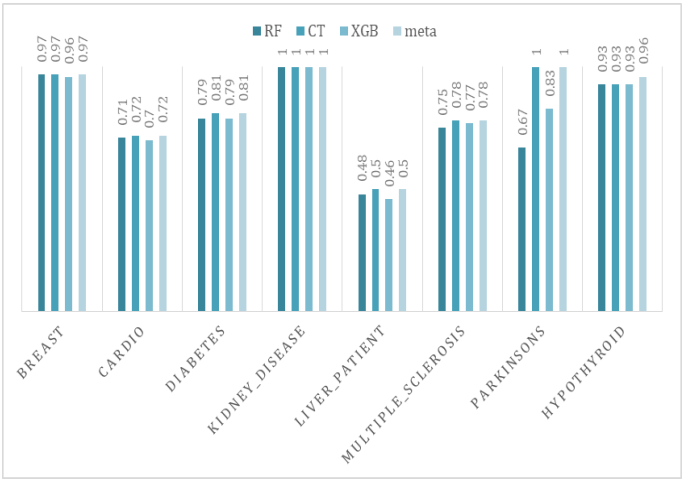

Fig. 3. Comparison based on precision

As show in **Fig 3.** for the Breast Dataset, the highest precision was 0.97 for both Random Forest and the meta model. As for the Cardio Dataset, the highest precision was 0.72 for the meta model. For the Diabetes Dataset, the highest precision is 0.81 for the CT and Meta models. In the case of the Kidney Disease Dataset, all models achieved a precision of 1.0, meaning they correctly classified all cases. In the Liver Patient Dataset, the highest precision is 0.5 for the CT and Meta models. Multiple Sclerosis Dataset The highest precision is 0.78 for the CT and Meta models and Parkinson's Dataset The highest precision is 1.0 for both the Random Forest and the Meta model. Finally, in the Hypothyroid Dataset, the highest precision is 0.96 for the meta model.

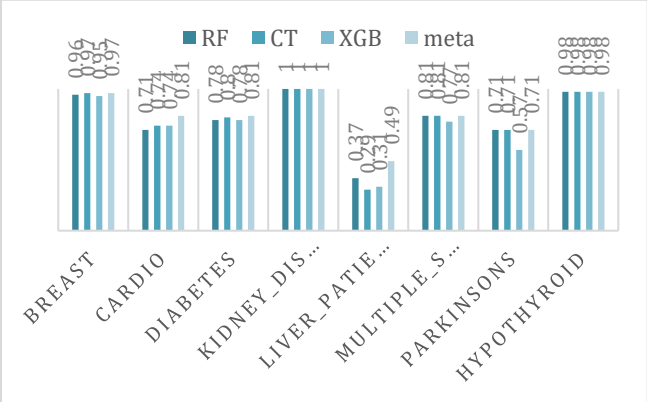

Fig. 4. Comparison based on Recall

Using the Recall evaluation as show in **Figure 4**, the meta model achieved the highest values in the case of the data sets (Breast, Cardio, Diabetes, Liver Patient, Multiple Sclerosis, Parkinson's, and Hypothyroid), which reached (0.97, 0.81, 0.81, 0.49, and 0.81). Dataset: The highest value for the meta model was 0.71. 0.98) respectively. While for Kidney Disease Dataset: All models got Recall 1.0, meaning they detected all positive cases correctly.

V. Discussions

The selected additive machine learning models, in addition to the proposed model, were tested on eight data sets for different diseases and the evaluation metrics (accuracy, precision, recall) were used. The results showed the superiority of the proposed meta-model, which indicates its effectiveness in classifying various diseases, as it was the highest when applying the model to the data sets (Breast, Diabetes, Liver Patient, Hypothyroidism, and Multiple Sclerosis), while it was close to the XGBoost model in the case of Parkinson's Dataset, and at the same time it was equal. Performance of the proposed model with Random Forest when applied to the kidney disease dataset.

Multiple datasets have shown high precision and recall results for the meta-model therefore signifying its potential to correctly diagnose positive cases without any errors. The relevance of this is especially important and critical in medical diagnosis where accuracy is important.

The meta-model has shown great success in accurately predicting disease outcomes. This has important implications for healthcare applications. It can help healthcare professionals detect diseases early, plan treatments, and make better decisions.

Researchers may explore the growing range of disease data sets in the future. They can look at how well the proposed meta-model works across different healthcare areas and how scalable it is. Other features like genetic markers or patient details could be added to improve the model's prediction ability, too. These additions may allow for more personalized healthcare interventions

VI. Conclusion

In this paper, a medical decision support system was developed and evaluated using a comprehensive multi-disease dataset that included breast cancer, cardiovascular health, diabetes, kidney diseases, liver patients, multiple sclerosis, Parkinson's disease, and thyroid conditions. The proposed approach used the stacking technique of ensemble learning algorithms to create a decision support system. Three machine learning models - CT, XGB and RF - are integrated, and a logistic regression model is used to combine the model outputs to create a decision support system. Three common performance indicators, precision, and recall—have also been used to measure the efficiency of models. In each scenario (disease), the meta-model achieved the best accuracy results. The proposed ensemble model obtained the highest accuracy, which is 97.36%, 73.70%, 75.32%, 100%, 70.69%, 82.01%, 94.87%, 99.58%, superior to other additive machine learning models. Future work could test the efficiency of the method with individual deep learning models.

Acknowledgement

The authors would express they're thanks to the college of computer science, University of Mosul, to support this report.

References

[1] Guo, C., & Chen, J. (2023). Big data analytics in healthcare. In Knowledge technology and systems: Toward establishing knowledge systems science (pp. 27-70). Singapore: Springer Nature Singapore.

[2] Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). Involvement of machine learning tools in healthcare decision making. Journal of healthcare engineering, 2021.

[3] Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O. A., Sepúlveda, M., Helm, E., ... & Zerbato, F. (2022). Process mining for healthcare: Characteristics and challenges. Journal of Biomedical Informatics, 127, 103994.

[4] Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F., & Tighe, P. (2021). Accessing artificial intelligence for clinical decision-making. Frontiers in digital health, 3, 645232.

[5] Buchlak, Q. D., Esmaili, N., Leveque, J. C., Farrokhi, F., Bennett, C., Piccardi, M., & Sethi, R. K. (2020). Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. Neurosurgical review, 43, 1235-1253.

[6] Selvaraj, S., & Sundaravaradhan, S. (2020). Challenges and opportunities in IoT healthcare systems: a systematic review. SN Applied Sciences, 2(1), 139.

[7] Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., ... & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. Clinical and translational science, 14(1), 86-93.

[8] Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks, 3, 58-73.

[9] Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. In Machine learning and the internet of medical things in healthcare (pp. 89-111). Academic Press.

[10] Kumar, N., Das, N. N., Gupta, D., Gupta, K., & Bindra, J. (2021). Efficient automated disease diagnosis using machine learning models. Journal of healthcare engineering, 2021.

[11] Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541). MDPI.

[12] Khan, P., Kader, M. F., Islam, S. R., Rahman, A. B., Kamal, M. S., Toha, M. U., & Kwak, K. S. (2021). Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. Ieee Access, 9, 37622-37655.

[13] Lohumi, P., Garg, S., Singh, T. P., & Gopal, M. (2020, October). Ensemble learning classification for medical diagnosis. In 2020 5th International Conference on Computing, Communication and Security (ICCCS) (pp. 1-5). IEEE.

[14] Mosavi, A., Sajedi Hosseini, F., Choubin, B., Goodarzi, M., Dineva, A. A., & Rafiei Sardooi, E. (2021). Ensemble boosting and bagging based machine learning models for groundwater potential prediction. Water Resources Management, 35, 23-37.

[15] Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. Neurocomputing, 510, 1-14.

[16] Konstantinov, A. V., & Utkin, L. V. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. Knowledge-Based Systems, 222, 106993.

[17] Koopialipoor, M., Asteris, P. G., Mohammed, A. S., Alexakis, D. E., Mamou, A., & Armaghani, D. J. (2022). Introducing stacking machine learning approaches for the prediction of rock deformation. Transportation Geotechnics, 34, 100756.

[18] Kalagotla, S. K., Gangashetty, S. V., & Giridhar, K. (2021). A novel stacking technique for prediction of diabetes. Computers in Biology and Medicine, 135, 104554.

[19] Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5 (pp. 108-117). Springer Singapore.

[20] Ma, F., Sun, T., Liu, L., & Jing, H. (2020). Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Generation Computer Systems, 111, 17-26.

[21] Assegie, T. A. (2021). An optimized K-Nearest Neighbor based breast cancer detection. Journal of Robotics and Control (JRC), 2(3), 115-118.

[22] Adamichou, C., Genitsaridi, I., Nikolopoulos, D., Nikoloudaki, M., Repa, A., Bortoluzzi, A., ... & Bertsias, G. K. (2021). Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. Annals of the rheumatic diseases, 80(6), 758-766.

[23] Ahmad, G.N.; Fatima, H.; Abbas, M.; Rahman, O.; Alqahtani, M.S. Mixed machine learning approach for efficient prediction of human heart disease by identifying the numerical and categorical features. Appl. Sci. 2022, 12, 7449.

[24] Ma, Y., Chen, J., Wang, T., Zhang, L., Xu, X., Qiu, Y., ... & Huang, W. (2022). Accurate machine learning model to diagnose chronic autoimmune diseases utilizing information from B cells and monocytes. Frontiers in Immunology, 13, 870531.

[25] Han, Y., Jin, Z., Ma, L., Wang, D., Zhu, Y., Chen, S., ... & Feng, X. (2023). Development of clinical decision models for the prediction of systemic lupus erythematosus and Sjogren's syndrome overlap. Journal of Clinical Medicine, 12(2), 535.

[26] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[27] Rajib Kumar Halder, November 10, 2020, "Cardiovascular Disease Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/7qm5-dz13.

[28] M. Kahn, "Diabetes," UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C5T59G.

[29] Rubini,L., Soundarapandian,P., and Eswaran,P.. (2015). Chronic Kidney Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C5G020.

[30] Ramana,Bendi and Venkateswarlu,N.. (2012). ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. https://doi.org/10.24432/C5D02C.

[31] Pineda, Benjamin; Flores Rivera, Jose De Jesus (2023), "Conversion predictors of Clinically Isolated Syndrome to Multiple Sclerosis in Mexican patients: a prospective study.", Mendeley Data, V1, doi: 10.17632/8wk5hjx7x2.1

[32] Little,Max. (2008). Parkinsons. UCI Machine Learning Repository. https://doi.org/10.24432/C59C74.

[33] Saket Sathe and Charu C. Aggarwal. LODES: Local Density meets Spectral Outlier Detection. SIAM Conference on Data Mining, 2016.