


Review Article

Boosting Learning Algorithms for Chronic Diseases Prediction: A Review

Israa Mohammed Hassoon 
Department of Mathematics
College of Science, Mustansiriyah University
Baghdad, Iraq
isrmo9@uomustansiriyah.edu.iq

ARTICLE INFO

Article History

Received: 27/06/2024

Accepted: 30/07/2024

Published: 30/09/2024

This is an open-access article under the CC BY 4.0 license:

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

Boosting algorithms are a set of machine learning techniques that are predicated on the notion that a weak learner's acquisition of multiple basic classifiers might yield results that are superior to those of any one simple classifier used alone. A comprehensive evaluation of regularly used boosting techniques against highly investigated diseases is lacking, despite the fact that boosting approaches have been used for disease prediction in many studies. Thus, the purpose of this work is to highlight the main algorithms and strategies in the boosting learning. The results of this work will help academics identify a more appropriate boosting approach to predict disease, as well as better understand current patterns and hotspots in diseases prediction models that use boosting learning. The results showed that adaboost algorithm outperformed other algorithms in terms of accuracy, achieving above 90%. This review also demonstrates how combining two boosting methods can increase the basic classifier's accuracy. By using AdaBoost and LightGBM, the accuracy reached 99.75%. XGBoost and Gradient Boosting techniques were employed more frequently in researches than other boosting algorithms.

Keywords: *Boosting Learning; Extreme Gradient Boosting; Gradient Boosting Algorithm; Adaptive Boosting; Light Gradient-Boosting; Disease Prediction.*

1. INTRODUCTION

Worldwide, machine learning algorithms are changing the way diseases are diagnosed, and boosting is a powerful method that has become popular because it can enhance model performance. Boosting is a machine learning technique that creates a strong ensemble model by merging the predictions of several weak models.

Boosting is a flexible way of placing several weak learners one after the other. The new approach makes a competent learner with less bias at the conclusion of the process by intuitively focusing on the findings that have been demonstrated to be the hardest to match up to this point [1]. Boosting is suitable for many machine learning applications, including regression, ranking, and classification, since it can manage complex data patterns, including non-linear correlations and interactions. Because boosting concentrates on misclassified data and assigns them more weights than other machine learning algorithms, it is less susceptible to noise in training dataset and lessens the influence of noisy data on the final outcome predictions [2].

Boosting approaches can still offer some understanding, nevertheless, by helping to comprehend the relative importance of different features in the process of prediction through feature significance rankings. Boosting is intended to help the model learn from its mistakes and enhance its performance iteratively by giving priority to samples that were misclassified in earlier iterations [3]. Identifying the disease that most closely matches a person's symptoms is known as disease diagnosis. The most difficult problem to diagnose is one of unclear symptoms and indicators; identifying the condition is essential to treating any illness. Based on historical training data, machine learning is a field that can assist in predicting the diagnosis of disease. Disease prediction has gained substantial attention as a study topic due to the abundance of available data. Researchers can use these databases to create disease prediction models for decision-making systems, which enables better illness diagnosis and treatment at an earlier stage. Early diagnosis and timely treatment are the

most effective ways to lower the rates of disease-related mortality.

To efficiently recognize a broad range of circumstances, numerous scientists have developed a variety of machine learning methods. A model that forecasts diseases and their therapies can be produced by machine learning techniques. A patient's health may be greatly impacted by common disorders such as skin cancer, kidney disease, diabetes, liver disease, heart disease and migraine disease. Improving the management of chronic pain requires figuring out the underlying causes of the condition and making it possible to customize treatment [6].

This study investigates diseases prediction studies based on boosting learning. Initially, a number of disease prediction models have been identified by searching through the research and analyzing the disease types that are taken into consideration. Finding key patterns in the boosting techniques applied to different base model learners, their accuracy, and the kinds of disease that have been researched in the literature are the goals of this research. Furthermore, a summary of the advantages and disadvantages of several boosting strategies is provided. The results of this study will aid in the establishment of research priorities by assisting researchers in better understanding current trends and hotspots in diseases prediction approaches that employ boosting learning.

In this review, the following research questions will be examined:

1. What techniques of boosting have been employed?
2. What patient's information database has been utilized?
3. What kind of diseases can boosting algorithms classified?

The objective is to add to the corpus of knowledge by addressing this questions and offering insight into how boosting approaches might be applied to enhance the management and treatment of patients.

An overview of the article's remaining sections is provided below: Section 2 provides a quick overview of the boosting learning. Section 3 presents details on the scientific studies based on boosting learning. Section 4 concluded this work.

2. BOOSTING LEARNING

A method for ensemble learning called "boosting" fits a dataset to progressively weaker learners. The goal of each successively fitted weak learner is to minimize the errors from the preceding one. New subsets are constructed from the items that the previous model misclassified. Then, by employing a cost function to integrate the weak models, the ensemble process enhances its performance. It clarified that, in contrast to bagging, each model operates independently prior to combining the inputs; there is no final model selection [5].

Boosting is a useful technique for addressing identification and regression issues. Figure 1 illustrates the flowchart of the boosting approach. In a machine learning ensemble learning, boosting makes the model easier to interpret and aids in lowering bias and variance. Boosting has the disadvantage that each classifier has to correct the mistakes made by its predecessors. The problem of scaling consecutive training in boosting represents numerous challenges in its implementation. As the number of repetitions increases, it becomes increasingly computationally expensive and susceptible to overfitting [7].

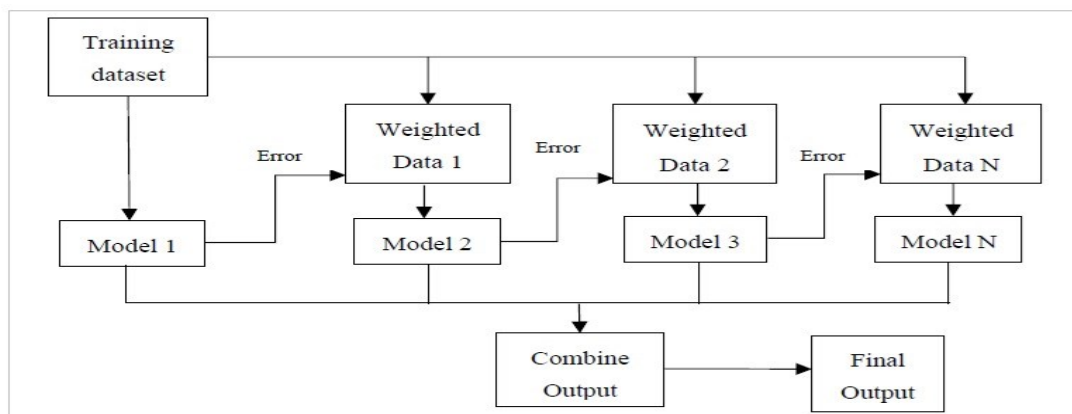


Fig. 1. The Framework Used in the Boosting Approach [8]

The boosting algorithm's steps [9]

- 1- Weight initialization: Every training example has the same weight at the beginning of the process.
- 2- Training: A Weak Learner is trained using the weighted training data. A weak learner is a basic model that very slightly beats random guessing.
- 3- Error computation: The weak learner's error on the training set is determined. The mistake is the weighted total of cases that were incorrectly classified.
- 4- Reset weights: Reset weights based on the training examples' error rate. Incorrectly classified instances have lower weights than correctly classified examples, with the former receiving higher weights.
- 5- Repeat steps 2-4. During each cycle, a fresh weak learner is educated using the newly adjusted weights of the training samples.
- 6- weak-learner combination: Every weak learner that was trained in the earlier stages is included in the final model. Every weak learner's accuracy is assigned a weight, and the final prediction is determined by adding up all of the weak learners' weights.
- 7- Predict: The completed model is utilized to estimate the class labels of new cases. Table 1 below illustrates the advantage and disadvantage of boosting learning algorithms.

TABLE I. THE ADVANTAGE AND DISADVANTAGE OF BOOSTING LEARNING ALGORITHMS

Boosting advantage	Boosting disadvantage
Boosting methods allow a model to provide a more accurate forecast that is unassailable.	Boosting methods have the potential to overemphasize the outliers and induce overfitting.
It works with both categorical and numeric variables, there is no need for data pre-processing.	The gradient boosting approach is computationally demanding because it continuously seeks to eliminate errors and necessitates several trees.
Compared to other algorithms, boosting techniques may optimize many loss functions and offer multiple hyperparameter tweaking choices, making them far more versatile.	It's a memory-intensive and time-consuming algorithm.
Missing data is handled automatically, therefore there is no need to impute in the dataset.	Several parameters are produced by these techniques' tremendous flexibility, which directly influences the model's behavior.
It's among the best methods for resolving two-class categorization problems.	Because of the increasing algorithmic complexity, real-time implementation of boosting is difficult.

2.1 BOOSTING LEARNING ALGORITHMS

2.1.1 GRADIENT BOOSTING MACHINE

Gradient Boosting Machine is a machine learning method that uses boosting in a functional space to produce a prediction model in the form of an ensemble of weak prediction models, or models that make minimal assumptions about the data and are usually straightforward decision trees. The algorithm that results when a decision tree is the weak learner is known as gradient-boosted trees. The gradient-boosted trees model is constructed in a stage-wise manner similar to other boosting methods, but it extends the capabilities of the other methods by permitting optimization of an arbitrary differentiable loss function. To obtain the final predictions, a GBM, aggregates the predictions from several decision trees. Recall that in a gradient-boosting machine, decision trees represent all of the weak learners. The Gradient Boosting method has been extensively investigated and utilized in numerous domains, including pharmacology, oil production monitoring, diesel engine fault detection, and hot spot classification [8, 10].

2.1.2 EXTREME GRADIENT BOOSTING (XGBOOST)

XGBoost is known as Extreme Gradient boosting, is another well-liked boosting technique. As a matter of fact, XGBoost is only the GBM algorithm modified, XGBoost functions according to the same process as GBM. In XGBoost, trees are constructed in a sequential manner with the goal of fixing the mistakes made by earlier trees. A key difference between XGBM and GBM is that the latter is slower due to its implementation of

parallel preprocessing. Numerous regularization strategies are also included in XGBoost, which lowers overfitting and boosts efficiency overall. Setting the XGBoost algorithm's hyperparameters will allow you to choose the regularization method [11]. XGBoost is a member of the gradient boosting framework, which is a subset of ensemble learning. It adds regularization techniques to improve the generalization of the model and uses decision trees as foundation learners. Widely utilized for tasks including regression, classification, and ranking, XGBoost is well-known for its computational efficiency, feature importance analysis, and management of missing values [12].

2.1.3 LightGBM

The gradient boosting technique LightGBM is tree-based and use leaf-wise tree growth rather than depth-wise development. Because of its speed and effectiveness, the LightGBM boosting algorithm is gaining popularity every day. Large volumes of data can be easily handled with LightGBM. However, bear in mind that fewer data points mean worse performance for this method. A framework for high-performance gradient boosting is called LightGBM. It is made to be accurate, scalable, and efficient. Its foundation is a decision tree architecture intended to lower memory consumption and increase model efficiency. Furthermore, LightGBM uses techniques based on histograms to generate trees more quickly. LightGBM's efficiency is enhanced by these methods, which also offer it an advantage over competing gradient boosting frameworks. Competing optimizations include leaf-wise tree development and effective data storage formats [13, 14].

2.1.4 AdaBoost

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an ensemble method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. It creates a model by assigning each data point an equal weight. Points that are incorrectly classified are therefore given greater weights. In the upcoming model, all the points with higher weights will now have greater significance. Until and unless a smaller error is received, it will continue training models [15].

AdaBoost uses a greedy approach to add the appropriately weighted predictor to the current model at each iteration in order to minimize the misclassification loss [16]. While Gradient Boost applies this strategy to any differential loss function, AdaBoost maximizes the exponential loss function.

AdaBoost approaches have been studied and compared with single classifier techniques like support vector machines and decision trees in the field of financial crisis prediction. Adaboost-based online algorithms are given for building local parameterized detection models on each node; this is another use of AdaBoost, namely dynamic distributed network intrusion detection [17].

3. RELATED WORK

This work reviewed a number of recent studies that predicted several serious diseases based on boosting learning. Boosting algorithms are employed to enhance the precision of forecasts generated by weaker models, like decision trees. The study in [18] examined and compared boosting algorithms on a diabetic dataset using knowledge discovery techniques. ROC curves are created and average accuracy values are compared in order to assess the efficacy of the boosting techniques. The study's results showed that the Gradient Boosting, CatBoost, AdaBoost, XGBoost, and LightGBM algorithms had success rates of %85, %88, %83, %87 and %86, when measured in terms of accuracy.

One well-liked ensemble learning technique, boosting, was used in [19] to increase CKD prediction accuracy. There are five included boosting algorithms. A machine learning repository at UCI called CKD provided data sets for their experiments. Together with appropriate hyperparameter tuning and feature selection, a number of preprocessing processes were used to improve prediction performance. Each characteristic in the dataset that contributed to CKD was evaluated for significance. Each model's performance was assessed using runtime, accuracy, recall, precision, F1-score, and ROC/AUC. AdaBoost was shown to perform the best overall, having the highest score across nearly all performance metrics. For the training and testing sets, it achieved 98.47% accuracy, respectively. Additionally, their model performed better on the AUC-ROC curve and showed improved recall, precision, and performance.

A gradient boosting method was used by DAS et al. [20] as the best performer in forecasting the probability of community death owing to COVID-19, demonstrating the usefulness of this algorithm in healthcare settings. In [21] an XGBoost-based diabetes prediction method is proposed, in which some significant characteristics

are derived from the experiment data's text features and the numerical features are separated. The findings of the experiment demonstrate that the enhanced XGBoost approach with feature combination has an accuracy of 80.2% for diabetes prediction, making it a practical and useful method for diabetes prediction.

In [22] a novel method that combines data mining classification techniques was introduced. They then created an ensemble approach that makes use of AdaBoost, and gradient boosting algorithms. To predict liver disease, Rehman, M.U. et al. [23] proposed the unique machine-learning approach to present a non-invasive iris feature-based method. Among the 879 patients from Pakistan in the experimental setup, 453 had chronic liver disease and 426 were healthy. This data set was used to train the models. With an infrared camera made up of a lens, a thermal sensor, and digital electronics processing, the iris images were gathered. By employing unique feature forms—22 physiological and 33 iris features—the lens concentrates on the infrared radiation on the sensor. Eleven separate classifiers were merged into a single classification framework for a non-invasive system, and the outcomes were compared using cross-validation techniques. Accuracy, F-score, precision, sensitivity, and specificity were the five parameters used to analyze the model's overall performance. These findings validated the non-invasive model's 98% prediction accuracy for chronic liver disorders.

In order to predict heart disease based on certain 12 medical parameters, Theerthagiri et al. [24] investigated a gradient boosting algorithm based on recursive feature elimination. In [25] a novel approach to get optimal features for individual models and their ensemble for skin disease was presented, combining four distinct filtering techniques with three integrated feature selection approaches. The suggested machine learning-based ensemble method was able to use the one vs. many classification strategy to categorize skin disease types into six groups on dermatology datasets. According to the findings, the Adaboost algorithm achieved 92.98 accuracy. Their study demonstrates how boosting learning techniques can more successfully and precisely forecast skin conditions.

In [26] a study proposed a prediction model for dementia risk using XGBoost and derived variable extraction from numerical dementia data. The method used gradient boosting to extract variable importance and generates derived variables. A Top-N group was created, and hyper-parameter tuning was conducted for optimal performance. The Top-20 model showed the best performance, with an accuracy of 85.61%.

In [27] a classifier based on the ensemble approach to enhance the effectiveness of kidney disease diagnosis classifiers was suggested. Through the use of ensemble methods, learning algorithms are combined to produce predicted performance that is superior to that of any one of the individual learning algorithms. Furthermore, the system's performance is measured on the receiver operational characteristic curve and data was examined using tenfold cross-validation. Their ensemble-based approach achieves the state-of-the-art performance, as demonstrated by extensive tests on CKD datasets from the UCI machine learning library.

In [28] the study used boosting algorithms (AdaBoost and XGBoost) to investigate Alzheimer's patients residing in South Korea in order to determine the predictors of anxiety. They also validated the machine learning algorithm with the best prediction performance. In order to conduct an early dementia screening in rehabilitation facilities, they examined 253 senior citizens with a diagnosis of Alzheimer's disease. Their study used XGBoost and AdaBoost to create models to predict the anxiety levels of people with Alzheimer's dementia. According to the study's findings, the model with the best prediction performance was found to be XGBoost based on SMOTE (accuracy=0.84, specificity= 0.81, and sensitivity=0.85). As a result, the results of the research indicated that employing an SMOTE-XGBoost model rather than an SMOTE-Adaboost model would result in greater accuracy.

In order to improve the accuracy of liver disease diagnosis, Adaboost and Firefly Algorithms were integrated in [29], ten machine learning features from the University of California, Irvine were included in the dataset, which has 583 independent records total. 20% of the data were used for testing and the remaining 80% for training. Compared to models without feature selection, the hybrid feature selection model demonstrated improved performance. The model's performance was impacted by the features chosen were 98.61% and 94.15%, respectively.

The study in [30] was investigated data-driven methods for identifying patients with these disorders by using supervised machine learning models. Their approach involves a comprehensive search of all feature variables present in the National Health and Nutrition Examination Survey (NHANES) dataset. This allowed them to create models that detect cardiovascular disease, pre-diabetes, and diabetes. Several machine learning models, including random forest, gradient boosting, logistic regression, and support vector machines, were tested for their ability to classify data using various time-frames and feature sets derived from laboratory analysis. The 131-variable ensemble model for cardiovascular disease that was built had an accuracy of 83.9% when laboratory results were used, and an (AU-ROC) score of 83.1% when laboratory values were not used.

EXtreme Gradient Boost (XGBoost) model got an AU-ROC score of 95.7% (with laboratory data) and 86.2% (without laboratory data) in diabetes prediction (based on 123 variables). With no laboratory data, the ensemble model got the highest AU-ROC score for pre-diabetic patients (73.7%), while XGBoost did best with laboratory-based data (84.4%).

A pipelined framework for diabetes prediction and classification was developed by A.Mujumdar and V. Vaidehi in [31]. A few important factors impacting the development of diabetes were taken into consideration, such as age, insulin, body mass index, glucose level, and so forth. The dataset consisted of eight hundred samples of patient data. Ten characteristics in all were picked for investigation. The implementation made use of several AI/ML strategies. The highest accuracy was 98% for AdaBoost.

In order to increase classification accuracy and promptly identify seizures, a machine learning approach based on a modified XGboost algorithm was used in [32]. To reduce sample mismatches between training and testing and improve classification model performance, the classic XGboost classifier model used a focused loss function. To evaluate the suggested classification model, the CHB-MIT SCALP Electroencephalography (EEG) dataset was used. That analysis of the suggested classification model's performance made use of the data collected for each of the 24 patients from the CHB-MIT Database. The proposed classification model's 2-class seizure trial outcomes were compared to a number of cutting-edge seizure classification models in this instance. The nature of the 2-class seizure was determined using cross-validation studies, where the prediction was seizure or non-seizure. The average specificity and sensitivity metrics scores were almost 100%. The suggested model outperforms the best conventional method in terms of average specificity at 1% and average sensitivity at 0.05%. The suggested modified XGBoost model performs better on average in terms of sensitivity and specificity than all current state-of-the-art seizure detection methods. Table-2 bellow illustrates more details about the reviewed studies dataset.

TABLE II. SUMMARY OF THE REVIEWED STUDIES DATASET

Ref. No.	Dataset Size	Dataset Type	Balancing Technique	No. of Healthy Instances	No. of Sick Instances	Dataset used
[17]	568	Balancing	SMOTE	500	268	Not mentioned
[18]	400	Balancing	SMOTE	150	250	Chronic Kidney Dataset collected from UCI machine learning repository
[19]	3,524	Balancing	SMOTE and ADASYN	Not mentioned	Not mentioned	Korea Center for Disease Control and Prevention
[20]	768	Unbalancing	Not used	Not mentioned	Not mentioned	National Institute of Diabetes and Digestive and Kidney Diseases
[21]	366	Unbalancing	Not used	112	254	University of California - Irvine machine learning repository (http://archive.ics.uci.edu/ml).
[22]	583	Unbalancing	Not used	416	167	UCI Indian Liver Patient
[23]	70000	Unbalancing	Not used	Not mentioned	Not mentioned	Kaggle repository
[24]	366	Unbalancing	Not used	112	254	UCI Dermatology
[26]	570	Unbalancing	Not used	Not mentioned	Not mentioned	Open Access Series of Imaging Studies (OASIS)

[27]	400	Unbalancing	Not used	250	150	UCI Chronic Kidney
[28]	253	Balancing	SMOTE	Not mentioned	Not mentioned	Central dementia center
[29]	400	Unbalancing	Not used	248	152	University of California, Irvine (UCI) repository
[30]	8459	Balancing	Under- sampling	7012	1447	National Health and Nutrition Examination Survey
[32]	983	Unbalancing	Not used	Not mentioned	Not mentioned	CHB-MIT Database

Table-3 bellow illustrates more details about studies based on the boosting algorithms.

TABLE III. SUMMARY OF REVIEWED STUDIES BASED ON BOOSTING ALGORITHMS

Ref. No.	Best Boosting Model	Features	Disease	Aim of Study	No. of classes	Accuracy (%)
[17]	CatBoost	9	Diabetes mellitus	Diagnosis	2	88%
[18]	AdaBoost	24	Kidney disease	Prediction	2	98.47%,
[19]	Gradient boosting	4	Covid-19	Prediction	2	97.1%
[20]	XGBoost Algorithm	8	Diabetes	Prediction	2	80.2%
[21]	Gradient Boosting	15	Skin disease	Prediction	6	99.46%
[22]	XGBoost	10	Liver Diseases	Diagnosis	2	86.7%
[23]	Gradient boosting	11	Cardiovascular	Prediction	2	89.7 %
[24]	AdaBoost	34	Erythematous Squamous	Prediction	6	97.4%
[26]	XGBoost	21	Dementia Risk	Prediction	2	85.61%
[27]	AdaBoost	25	Kidney Disease	Prediction	2	99%
[28]	XGBoost	7	Alzheimer's Dementia	Prediction	2	84%
[29]	AdaBoost and LightGBM	25	Kidney Disease	Prediction	2	99.75%

[30]	XGBoost	131	Diabetes and Cardiovascular Disease	Prediction	2	83.8%
[32]	XGBoost	24	Seizure	Prediction	2	99.99%

4. CONCLUSIONS

Boosting learning algorithms are the most commonly used to develop and improve various early disease prediction systems because they lower bias and variance. This literature review discusses that boosting approaches compared to other machine learning methods increase classifiers' accuracy. This work offers a thorough analysis of the studies on chronic diseases prediction models that make use of different boosting types, along with a summary of the boosting methodology. The literature-based publications' descriptions offer crucial details on the performance of these algorithms in different configurations. Additionally, this study can help researchers to identify the best boosting technique for disease forecasting. The review's findings demonstrated that in comparison to other algorithms, the adaboost algorithm attained a high accuracy of over 90%. The using of two boosting algorithms together can increase accuracy. By using AdaBoost and LightGBM, the accuracy was 99.75%. XGBoost and Gradient Boosting techniques were employed more frequently in studies than other boosting algorithms.

Acknowledgment

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

REFERENCES

- [1] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>.
- [2] E. K. Sahin, “Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest,” *SN Applied Sciences*, vol. 2, no. 7, p. 1308, 2020. .
- [3] K. Qadeer, and M. Jeon, (2019). Prediction of PM10 concentration in South Korea using gradient tree boosting models. *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*. <https://doi.org/10.1145/3387168.3387234>.
- [4] E.C. Igodan, A.F.-B. Thompson, O. Obe, O Owolafe Erythemato Squamous, Disease Prediction using Ensemble Multi-Feature Selection Approach. *Int. J. Comput. Sci. Inf. Secur. IJCSIS* **2022**, 20, 95–106.
- [5] M. Khayamnia, M. Yazdchi, A. Heidari, and M. Foroughipour, (2019). Diagnosis of common headaches using hybrid expert-based systems. *Journal of Medical Signals and Sensors*, 9(3), 174. <https://doi.org/10.4103/jmss.jmss.47.18>.
- [6] S. Kumari; D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int. J. Cogn. Comput. Eng.* **2021**, 2, 40–46.
- [7] Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University. Computer and Information Sciences*, 35(2), 757–774. <https://doi.org/10.1016/j.jksuci.2023.01.014>.
- [8] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A review. *Healthcare*, 11(12), 1808. <https://doi.org/10.3390/healthcare11121808>.
- [9] Schapire R.(1999). The strength of weak learnability. *Machine Learning*, 5, 197–227.
- [10] Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., & Javed, T. (2022). Cardiovascular Disease Detection using Ensemble Learning. *Computational Intelligence and Neuroscience*, 2022, 1–9. <https://doi.org/10.1155/2022/5267498>.
- [11] Shin, Y. (2019). Application of stochastic gradient boosting approach to early prediction of safety accidents at construction site. *Advances in Civil Engineering*, 2019, 1–9. <https://doi.org/10.1155/2019/1574297>.
- [12] Qawasmeh, A., Alhusan, N., Hanandeh, F., & Al-Atiyat, M. (2020). A high performance system for the diagnosis of headache via hybrid machine learning model. *International Journal of Advanced Computer Science and Applications*, 11(5). <https://doi.org/10.14569/ijacsa.2020.0110580>.

- [13] Huang, K. (2020). An optimized LightGBM model for fraud detection. *Journal of Physics. Conference Series*, 1651(1), 012111. <https://doi.org/10.1088/1742-6596/1651/1/012111>.
- [14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017) LightGBM: a highly efficient gradient boosting decision tree. In: 2017 Conference on Neural Information Processing Systems. Long Beach, CA, USA. pp. 3149-3157.
- [15] Zakariah, M., AlQahtani, S. A., & Al-Rakhami, M. S. (2023). Machine Learning-Based Adaptive Synthetic Sampling technique for intrusion detection. *Applied Sciences*, 13(11), 6504. <https://doi.org/10.3390/app13116504>.
- [16] Ramesh, D.; Katheria, Y.S. Ensemble method based predictive model for analyzing disease datasets: A predictive analysis approach. *Health Technol.* **2019**, 9, 533–545.
- [17] Koçak, H., & ÇetiN, G. (2023). The Diagnosis of Diabetes Mellitus with Boosting Methods. *El-cezeri*. <https://doi.org/10.31202/ecjse.1242207>.
- [18] Ganie SM, Dutta Pramanik PK, Mallik S, Zhao Z. Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PLoS One*. 2023 Dec 1;18(12):e0295234. doi: [10.1371/journal.pone.0295234](https://doi.org/10.1371/journal.pone.0295234). PMID: 38039306; PMCID: PMC10691694.
- [19] Das, A. K., Mishra, S., & Gopalan, S. S. (2020). Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ*, 8, e10083. <https://doi.org/10.7717/peerj.10083>.
- [20] Li, M., Fu, X., & Li, D. (2020). Diabetes prediction based on XGBOOST algorithm. *IOP Conference Series. Materials Science and Engineering*, 768(7), 072093. <https://doi.org/10.1088/1757-899x/768/7/072093>.
- [21] Verma, A.K.; Pal, S.; Kumar, S. and Tiwan, B.B. Comparison of skin disease prediction by feature selection using ensemble data mining techniques. *Inform. Med. Unlocked* **2019**, 16, 100202.
- [22] Singh, V.; Gourisaria, M.K.; Das, H. Performance Analysis of Machine Learning Algorithms for Prediction of Liver Disease. In Proceedings of the 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCon), Kuala Lumpur, Malaysia, 24–26 September 2021; pp. 1–7.
- [23] P. Theerthagiri and J. Vidya, “Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques,” *CoRR*, vol. abs/2106.0, 2021. [Online]. Available: <https://arxiv.org/abs/210608889>.
- [24] Igodan, E.C.; Thompson, A.F.-B.; Obe, O.; Owolafe, O. Erythematous Squamous. (2022). Disease Prediction using Ensemble Multi-Feature Selection Approach. *Int. J. Comput. Sci. Secur. IJCSIS*, 20, 95–106.
- [25] K. Budholiya, S. K. Shrivastava and V. Sharma, “An optimized XGBoost based diagnostic system for effective prediction of heart disease,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, 2022.
- [26] Ryu, S., Shin, D., & Chung, K. (2020). Prediction model of dementia risk based on XGBOOST using derived variable extraction and hyper parameter optimization. *IEEE Access*, 8, 177708–177720. <https://doi.org/10.1109/access.2020.3025553>.
- [27] Hasan Z.K.M., Hasan, Md. Z. (2019). Performance evaluation of ensemble-based machine learning techniques for prediction of chronic Kidney disease. In Emerging Research in Computing, Information, Communication and Applications; Springer: Berlin/Heidelberg, Germany, pp. 415–426.
- [28] Byeon, H. (2021). Predicting the Anxiety of Patients with Alzheimer’s Dementia using Boosting Algorithm and Data-Level Approach. *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications*, 12(3). <https://doi.org/10.14569/ijacsa.2021.0120313>.
- [29] Sheyda Ardham, and farhad soleimanian gharehchopogh, “diagnosing liver disease using firefly algorithm based on adaboost,” *journal of health administration*, vol. 22, no. 1 (75), pp. 61–77, 2019, [Online]. Available: <https://sid.ir/paper/361212/en>.
- [30] Dinh, A.; Miertschin, S.; Young, A.; Mohanty, S.D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with Machine learning. *BMC Med. Inf. Decis. Mak.* 19, 211.
- [31] Mujumdar A. and Vaidehi V. (2019). Diabetes Prediction using Machine Learning Algorithms, *Procedia Comput. Sci.*, vol. 165, pp. 292–299, doi: [10.1016/j.procs.2020.01.047](https://doi.org/10.1016/j.procs.2020.01.047).
- [32] Kumar, T. H. R., Narayanappa, C. K., Raghavendra, S., & Poornima, G. R. (2022). A modified XG Boost classifier model for detection of Seizures and Non-Seizures. *WSEAS Transactions on Biology and Biomedicine*, 19, 14–21. <https://doi.org/10.37394/23208.2022.19.3>.