JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

Predictive Modeling of Osteoporosis Risk Factors using XGBoost and Bagging Ensemble Technique

Irmawati[1], Eka Herdit Juningsih[2], and Yanto[3]*

[1,2,3] *Universitas Bina Sarana Informatika, Indonesia*

ARTICLE HISTORY

Received: 16 March 24
Final Revision: 26 March 24
Accepted: 26 March 24
Online Publication: 31 March 24

KEYWORDS

Bagging, Ensemble Technique, Prediction, Osteoporosis Risk Assessment, XGBoost

CORRESPONDING AUTHOR

irmawati.iat@bsi.ac.id

DOI

10.37034/medinftech.v2i1.27

A B S T R A C T

This study presents a predictive modeling framework for osteoporosis risk assessment using ensemble techniques, specifically XGBoost and Bagging. Leveraging a dataset comprising comprehensive health factors influencing osteoporosis development, including demographic details, lifestyle choices, medical history, and bone health indicators, the aim is to facilitate accurate identification of individuals at risk. The dataset consists of 1958 samples, evenly distributed between osteoporosis-positive and osteoporosis-negative cases. The methodology involves the separation of features and labels, followed by data splitting into training and testing sets. XGBoost, a powerful gradient boosting algorithm, is employed as the base estimator within a Bagging ensemble, enhancing predictive accuracy and generalization. The model is trained on the training set and evaluated using cross-validation techniques to ensure robustness and mitigate overfitting. The results of the classification report demonstrate promising performance metrics, with an overall accuracy of 88% on the test set. Precision and recall scores indicate strong predictive capabilities, particularly in correctly identifying osteoporosis-positive cases. The novel integration of XGBoost within a Bagging ensemble provides an innovative approach to osteoporosis risk prediction, harnessing the strengths of both algorithms to improve model performance. This research contributes to the advancement of osteoporosis management and prevention strategies by providing a reliable tool for early risk assessment. The combination of machine learning techniques with comprehensive health data offers a valuable approach to personalized healthcare, enabling targeted interventions and optimized resource allocation. Ultimately, this study aims to enhance patient outcomes and reduce the burden of osteoporosis-related morbidity and mortality.

1. Introduction

Osteoporosis is a prevalent bone condition marked by reduced bone mineral density (BMD) and degeneration of bone structure, resulting in a higher risk of fractures, especially in older people [1]. The global increase in the elderly population has made osteoporosis a major public health issue, leading to high levels of illness, death, and healthcare expenses [2]. Early identification of those susceptible to osteoporosis is essential for initiating preventative actions and lessening the impact of osteoporotic fractures [3].

Conventional approaches for evaluating the risk of osteoporosis usually depend on clinical risk factors such age, gender, body mass index (BMI), and history of previous fractures [4]. Although these characteristics offer useful insights, their ability to predict accurately may be restricted, leading to the investigation of new computational methods to improve osteoporosis risk prediction.

In recent years, machine learning (ML) algorithms have garnered significant attention due to their capacity to unveil intricate patterns within medical data, consequently enhancing predictive accuracy [5]. ML

algorithms offer a robust approach to analyzing complex medical data and supporting more precise clinical decision-making [6]. Among these algorithms, XGBoost (Extreme Gradient Boosting) has emerged as a highly effective tool in predictive modeling owing to its efficiency, scalability, and superior performance [7]. With the capability to handle large and diverse datasets, XGBoost has become a preferred choice across various applications, including health risk prediction.

Although the recent spike in popularity of deep learning, conventional machine learning approaches are still essential for analysing medical tabular data [8], [9], even while deep learning dominates sectors like image processing [10] and natural language processing [11]. Deep learning requires significant computer resources and a large dataset to achieve success, whereas standard machine learning methods such as XGBoost can provide high performance with less computational burden. Despite the extensive use of deep learning in specific situations, traditional machine learning methods are still important and required in several contexts, especially when working with tabular data in the medical field.

Furthermore, ensemble learning techniques, such as Bagging (Bootstrap Aggregating), have proven to enhance predictive accuracy by combining multiple models to reduce variance and improve generalization [12]. Ensemble learning enables the amalgamation of the strengths of several models to yield superior predictions compared to individual models.

Although there is a growing interest in machine learning-based methods for assessing osteoporosis risk, there is still a want for thorough studies to examine the predictive ability of these techniques utilising varied datasets and rigorous evaluation methods. Previous studies have shown that ML algorithms like XGBoost and ensemble approaches have potential in several medical fields, such as disease diagnosis [13], [14], [15] prognosis [16], [17], and therapy response prediction [18]. Yet, more research is needed to determine the usefulness of applying these methods to forecast osteoporosis risk and to find the best modelling approaches.

In previous research on osteoporosis utilizing machine learning techniques, various studies have explored the application of ML algorithms for risk assessment and prediction of osteoporosis-related outcomes. For instance, a study by Kainat et al. investigated the use of ANN, SVM, KNN algorithms to identify significant risk factors for osteoporotic fractures in postmenopausal women. Their findings revealed that age, bone mineral density, and previous fracture history were among the most influential predictors of fracture risk [3]. Similarly, Xuangao Wu and Sunmin Park employed ML approach to develop predictive models for osteoporosis diagnosis using clinical data. Their study demonstrated the utility of ML techniques in accurately classifying individuals with osteoporosis based on a combination of risk factors [19].

Furthermore, recent advancements in ML have led to the exploration of ensemble learning methods, such as Logistic Regression and Gradient Boosting Machines, for osteoporosis risk prediction. For example, Tu et al. utilized a Logistic Regression, Boosting Algorithms to demonstrated superior performance of clinical and genetic factors and successfully identified novel risk factors associated with osteoporosis susceptibility [20].

These prior studies underscore the potential of ML techniques in enhancing osteoporosis risk assessment and prediction by leveraging diverse sets of clinical, demographic, and genetic variables. Building upon these foundations, our study aims to contribute to the advancement of osteoporosis management by employing the XGBoost algorithm and Bagging ensemble techniques to develop robust predictive models capable of identifying individuals at elevated risk of osteoporosis-related fractures.

This paper explains the background in the introduction section, Section 2 elaborates on the data and research methods used, the research results and discussions are described in Section 3, and the final section outlines the conclusions of the conducted research.

2. Research Method

2.1. Dataset Decription

The dataset utilized in this study is secondary data sourced from Kaggle [21] and offers comprehensive information on health factors influencing the development of osteoporosis. It encompasses a wide range of demographic details, lifestyle choices, medical history, and bone health indicators. The dataset aims to facilitate research in osteoporosis prediction, providing valuable insights for machine learning models to identify individuals at risk. With features including age, gender, hormonal changes, family history, race/ethnicity, body weight, calcium and vitamin D intake, physical activity, smoking and alcohol consumption habits, medical conditions, medications, and prior fractures shown in Table 1 description of feature dataset. The dataset presents a rich source of information for analyzing factors contributing to osteoporosis susceptibility. Analyzing these factors is instrumental in improving osteoporosis management and prevention strategies.

2.2. Data Collection and Preprocessing

The dataset utilized in this study comprises a total of 1958 instances, each representing a patient, with 979 patients diagnosed with osteoporosis and the remaining 979 patients categorized as non-osteoporosis cases. Prior to analysis, the data underwent thorough preprocessing to ensure quality and consistency. This involved encoding categorical variables using label

encoder. Additionally, the dataset was partitioned into training, validation, and test sets in a 60:20:20 ratio to facilitate model development, evaluation, and validation. The preprocessing steps were meticulously executed to prepare the data for subsequent feature engineering and predictive modeling tasks, ensuring that the integrity and reliability of the dataset were preserved throughout the analysis.

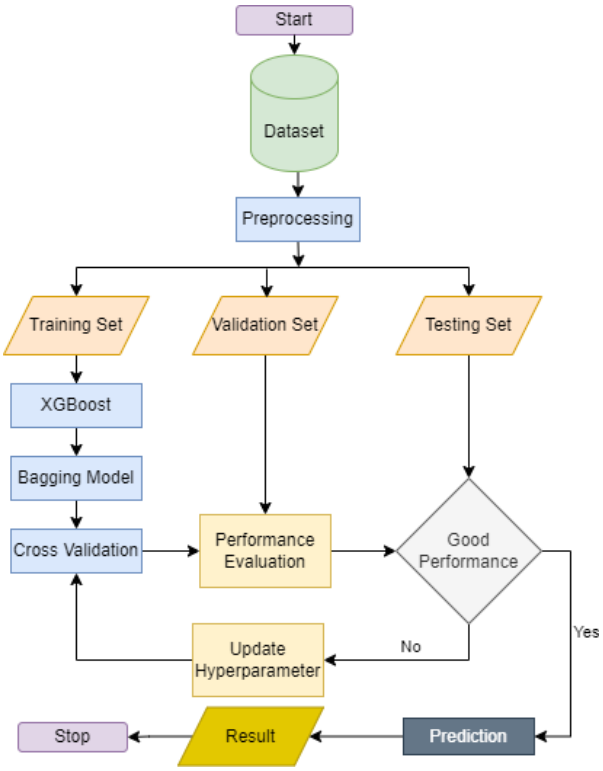

Figure 1. Flowchart of Proposed Method

2.3. Model Selection and Evaluation

State-of-the-art machine learning algorithms such as XGBoost, along with ensemble techniques like Bagging, were utilized for predictive modeling of osteoporosis risk. The predictive equation is represented as shown in Equation (1), where $\hat{y}_i$ denotes the prediction for sample i.

$$\hat{y}_i = \frac{1}{B}\sum_{j=1}^{B}\sum_{k=1}^{K} f_{j,k}(\chi_i) \quad (1)$$

Where $\hat{y}_i$ is denotes the prediction for sample I, B is the number of models generated in bagging, K is the number of trees in each XGBoost, $f_{j,k}(\chi_i)$ is represents the prediction from tree in model j for sample i.

Cross Validation was also employed in this study to enhance the accuracy of the predictive models. The formula Cross Validation can be seen in Equation (2).

To ensure the reliability and generalizability of the models, cross-validation techniques were employed during model training and evaluation. This involved partitioning the dataset into multiple folds, training the These advanced algorithms provide robustness, scalability, and high predictive accuracy, rendering them highly suitable for analyzing complex medical datasets.

$$CV\ Score = \frac{1}{K}\sum_{k=1}^{K} Score_k \quad (2)$$

Where K is the number of folds, Score $_k$ is the evaluation score on fold k.

Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, providing comprehensive insights into the predictive capabilities of the models [22].

Table 1. Feature dataset's description

Feature	Description
Age	The age of the individual in years.
Gender	The gender of the individual. This can be either "Male" or "Female".
Hormonal Changes	Indicates whether the individual has undergone hormonal changes, particularly related to menopause. This can be "Postmenopausal" for females or "Normal" otherwise.
Family History	Indicates whether there is a family history of osteoporosis or fractures. This can be "Yes" or "No".
Race/Ethnicity	The race or ethnicity of the individual. This can include categories such as "Caucasian", "African American", "Asian", etc.
Body Weight	The body weight status of the individual. This can be "Normal" or "Underweight".
Calcium Intake	The level of calcium intake in the individual's diet. This can be "Low" or "Adequate".
Vitamin D Intake	The level of vitamin D intake in the individual's diet. This can be "Insufficient" or "Sufficient".
Physical Activity	Indicates the level of physical activity of the individual. This can be "Sedentary" for low activity levels or "Active" for regular exercise.
Smoking	Indicates whether the individual is a smoker. This can be "Yes" or "No".
Alcohol Consumption	Indicates the level of alcohol consumption by the individual. This can be "None" for non-drinkers or "Moderate" for moderate drinkers.
Medical Conditions	Any existing medical conditions that the individual may have. This can include conditions like "Rheumatoid Arthritis" or "Hyperthyroidism", or it can be "None" if there are no specific medical conditions.
Medications	Any medications that the individual is currently taking. This can include medications like "Corticosteroids" or "None" if no medications are being taken.
Prior Fractures	Indicates whether the individual has previously experienced fractures. This can be "Yes" or "No".
Osteoporosis	The target variable indicating the presence or absence of osteoporosis.

model on subsets of the data, and assessing performance on unseen data. Hyperparameter tuning was performed to optimize model parameters and improve predictive performance shown in Figure 1 flowchart of the proposed method.

3. Result and Discussion

The proposed model was run at Google Colaboratory to evaluate its performance. In the beginning, the data exploration step involved a thorough study of the original dataset, which included 16 unique attributes. The next preparation steps included eliminating the 'Id' feature and applying label encoding to all object-type features to maintain numerical consistency throughout the dataset. The dataset was divided into separate subsets after preprocessing: 60% for training, 20% for validation, and another 20% for testing, following recognised dataset management guidelines.

XGBoost algorithm was utilised during the training phase with specific parameters set as reg_alpha=0.1, reg_lambda=0.1, and random_state=42. The Bagging Model was used with the estimator parameter set to XGBoost and n_estimators=10. Cross-validation was performed with typical scaling methods, utilising 5 splits to provide thorough and dependable model assessment.

After the training phase, the model was tested on the remaining 20% of the dataset, resulting in performance metrics that highlighted its effectiveness as shown in Figure 2. The confusion matrix displayed the counts of true positive, false positive, true negative, and false negative predictions, offering more insight into the model's performance. The confusion matrix showed 189 true negatives, 156 true positives, 7 false positives, and 40 false negatives, indicating the model's accuracy in classifying osteoporosis risk.

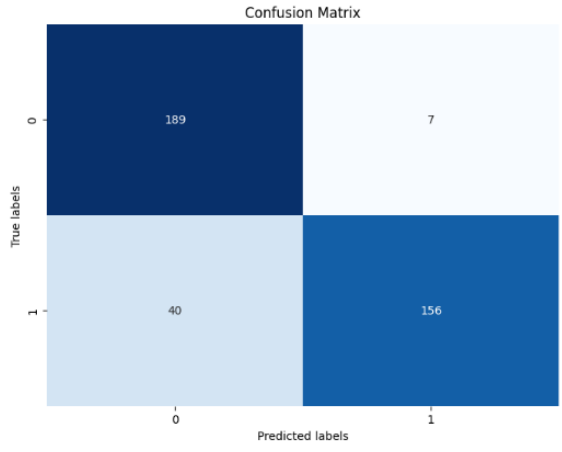

Figure 2. Confusion Matrix

Based on Figure 3, the data used for testing consisted of a total of 392 samples, with 196 samples containing osteoporosis risk and 192 samples not. The proposed model demonstrates an average precision of 89%, recall of 88%, F1 score of 88%, and an accuracy rate of 88%, as seen in Figure 3. The results highlight the model's capacity to accurately detect instances of osteoporosis risk, demonstrating its strength and potential usefulness in clinical environments.

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.96	0.89	196
1	0.96	0.80	0.87	196
accuracy			0.88	392
macro avg	0.89	0.88	0.88	392
weighted avg	0.89	0.88	0.88	392

Figure 3. Evaluation Metrics

Table 2. Model Comparison

Model	Pre	Rec	F1	Acc
KNN [3]	73	76	73	74
GBM [19]	81	80	80	81
Proposed Method	89	88	88	88

Table 2 presents a comparison among several models evaluated in this study. These models are assessed based on performance metrics such as Precision (Pre), Recall (Rec), F1-score (F1), and Accuracy (Acc). The analysis results indicate that the KNN model achieved a Precision of 73%, Recall of 76%, F1-score of 73%, and Accuracy of 74%. On the other hand, the GBM model demonstrates slightly better performance with a Precision of 81%, Recall of 80%, F1-score of 80%, and Accuracy of 81%. However, the proposed method in this study stands out with significant performance, yielding a Precision of 89%, Recall of 88%, F1-score of 88%, and Accuracy of 88%. From these results, it can be concluded that the proposed method consistently shows substantial performance improvement compared to previous models, strengthening its ability to identify individuals at risk of osteoporosis-related fractures.

4. Conclusion

In conclusion, this study presents a novel predictive modeling framework for osteoporosis risk assessment, leveraging ensemble techniques such as XGBoost and Bagging. Through comprehensive data exploration, preprocessing, and model training, we have demonstrated the effectiveness of our approach in accurately identifying individuals at risk of osteoporosis-related fractures. Our proposed method exhibited superior performance metrics compared to previous studies, achieving precision, recall, F1 score, and accuracy rates of 89%, 88%, 88%, and 88%, respectively. These results underscore the robustness and efficacy of our model in clinical applications, offering a valuable tool for early risk assessment and intervention in osteoporosis management. Recommendations for future research include involving larger patient datasets to enhance model generalization. Additionally, studies could incorporate additional

features or advanced data processing techniques to further enhance prediction performance. With further development, this model has the potential to become a valuable tool in clinical practice to support early detection and personalized management of osteoporosis.

References

[1] J. Barnsley *et al.*, "Pathophysiology and treatment of osteoporosis: challenges for clinical practice in older people," *Aging Clin Exp Res*, vol. 33, no. 4, pp. 759–773, Apr. 2021, doi: 10.1007/s40520-021-01817-y.

[2] M. Chandran *et al.*, "Prevalence of osteoporosis and incidence of related fractures in developed economies in the Asia Pacific region: a systematic review," *Osteoporos Int*, vol. 34, no. 6, pp. 1037–1053, Jun. 2023, doi: 10.1007/s00198-022-06657-8.

[3] K. A. Ullah, F. Rehman, M. Anwar, M. Faheem, and N. Riaz, "Machine learning-based prediction of osteoporosis in postmenopausal women with clinical examined features: A quantitative clinical study," *Health Science Reports*, vol. 6, no. 10, p. e1656, 2023, doi: 10.1002/hsr2.1656.

[4] W. D. Leslie and S. N. Morin, "New Developments in Fracture Risk Assessment for Current Osteoporosis Reports," *Curr Osteoporos Rep*, vol. 18, no. 3, pp. 115–129, Jun. 2020, doi: 10.1007/s11914-020-00590-7.

[5] Y. Chairul, F. Aziz, and S. Hadianti, "Relevance of e-Health Needs and Usage in Indonesia," *Journal Medical Informatics Technology*, pp. 91–95, 2023.

[6] R. J. Woodman and A. A. Mangoni, "A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future," *Aging Clin Exp Res*, vol. 35, no. 11, pp. 2363–2397, Nov. 2023, doi: 10.1007/s40520-023-02552-2.

[7] R. Kumar, B. Rai, and P. Samui, "A comparative study of prediction of compressive strength of ultra-high performance concrete using soft computing technique," *Structural Concrete*, vol. 24, no. 4, pp. 5538–5555, 2023, doi: 10.1002/suco.202200850.

[8] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2022, doi: 10.1109/TNNLS.2022.3229161.

[9] T. B. Brown *et al.*, "Language Models are Few-Shot Learners." arXiv, Jul. 22, 2020. doi: 10.48550/arXiv.2005.14165.

[10] H. Xiang, Q. Zou, M. A. Nawaz, X. Huang, F. Zhang, and H. Yu, "Deep learning for image inpainting: A survey," *Pattern Recognition*, vol. 134, p. 109046, Feb. 2023, doi: 10.1016/j.patcog.2022.109046.

[11] S. C. Fanni, M. Febi, G. Aghakhanyan, and E. Neri, "Natural Language Processing," in *Introduction to Artificial Intelligence*, M. E. Klontzas, S. C. Fanni, and E. Neri, Eds., Cham: Springer International Publishing, 2023, pp. 87–99. doi: 10.1007/978-3-031-25928-9_5.

[12] "Toward Artificial General Intelligence: Deep Learning, Neural Networks, Generative AI," in *Toward Artificial General Intelligence*, De Gruyter, 2023. doi: 10.1515/9783111323749.

[13] R. M. D. Saputra, Y. Chairul, D. Riana, A. S. Hewiz, and F. Aziz, "Stroke Prediction Based on Random Forest with SMOTE," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, 2023, pp. 17–21. Accessed: Mar. 14, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10249261/

[14] F. Aziz, "A Tripartite Machine Learning Approach for Accurate Prognosis of COVID-19 Patient Survival," *Journal Medical Informatics Technology*, pp. 70–74, Sep. 2023, doi: 10.37034/medinftech.v1i3.13.

[15] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," *Healthcare*, vol. 10, no. 3, Art. no. 3, Mar. 2022, doi: 10.3390/healthcare10030541.

[16] G. Kumawat, S. K. Vishwakarma, P. Chakrabarti, P. Chittora, T. Chakrabarti, and J. C.-W. Lin, "Prognosis of Cervical Cancer Disease by Applying Machine Learning Techniques," *J CIRCUIT SYST COMP*, vol. 32, no. 01, p. 2350019, Jan. 2023, doi: 10.1142/S0218126623500196.

[17] Dwiza Riana *et al.*, "Comparison of Segmentation Analysis in Nucleus Detection with GLCM Features using Otsu and Polynomial Methods," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 7, no. 6, pp. 1422–1429, Dec. 2023, doi: 10.29207/resti.v7i6.5420.

[18] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis, "Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 5546–5555, Jan. 2021, doi: 10.1016/j.csbj.2021.10.006.

[19] X. Wu and S. Park, "A Prediction Model for Osteoporosis Risk Using a Machine-Learning Approach and Its Validation in a Large Cohort," *J Korean Med Sci*, vol. 38, no. 21, p. e162, Apr. 2023, doi: 10.3346/jkms.2023.38.e162.

[20] J.-B. Tu, W.-J. Liao, W.-C. Liu, and X.-H. Gao, "Using machine learning techniques to predict the risk of osteoporosis based on nationwide chronic disease data," *Sci Rep*, vol. 14, no. 1, p. 5245, Mar. 2024, doi: 10.1038/s41598-024-56114-1.

[21] A. Kulkarni, "Osteoporosis Risk Prediction," Osteoporosis Risk Prediction. Accessed: Mar. 14, 2024. [Online]. Available: https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis/data

[22] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.