Intelligent Information Management, 2025, 17(2), 31-47
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

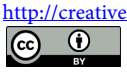
Enhancing Prediction of Osteoporosis Using Supervised and Unsupervised Learning: New Approach to Disease Subtyping

Muhannad Almohaimeed

Department of Information Systems, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia
Email: mmohimeed@taibahu.edu.sa

How to cite this paper: Almohaimeed, M. (2025) Enhancing Prediction of Osteoporosis Using Supervised and Unsupervised Learning: New Approach to Disease Subtyping. *Intelligent Information Management*, **17**, 31-47.
https://doi.org/10.4236/iim.2025.172002

Received: February 11, 2025
Accepted: March 16, 2025
Published: March 19, 2025



Abstract

Background: Osteoporosis is a serious health issue that can lead to severe clinical diseases, involving fractures. The resulting fracture can be a major risk factor for disability or even death for the elderly. A well-timed diagnosis of osteoporosis disease can help identify and prevent such fractures and improve patient outcomes. **Objective:** The aim of this study is to explore a novel hybrid approach for the characterization of osteoporotic patients into different subtypes, leading to enhanced classification of the condition. **Methods:** We examined a cohort of 10,000 patients based on nationwide chronic disease data in Germany, which included 1293 osteoporotic patients. We included various medical variables such as chronic kidney disease, cancer, stroke, hypertension, and diabetes. We deployed a hybrid approach that used HDBSCAN clustering to stratify patients into distinct subtypes. We constructed the predictive models for each subtype using seven different classification methods. **Results:** We identified seven distinct subtypes, each linked with different conditions such as cancer, cardiovascular diseases, and chronic obstructive pulmonary disease (COPD). Logistic Regression showed the highest subtype-level prediction performance, reaching an accuracy score of 87.8% compared to other predictive models based on the original dataset without clustering. Unsupervised learning approach improved prediction using all classification methods, emphasizing the impact of deploying subtype analysis to complex data. **Conclusion:** This research revealed that deploying a hybrid methodology is important for the discovery of patient subtypes and for making predictions more precise. The choice of the methods in this research was critical in ensuring robust prediction performance. The predictive model is vital for finding patients at high risk for osteoporosis disease and enabling early intervention and prevention strategies. This approach holds potential for the study of other complex clinical diseases using any data source.

Keywords

Osteoporosis, Machine Learning, Prediction, Clustering, Subtypes

1. Introduction

Osteoporosis is one of the major public health issues worldwide, which can lead to significant medical consequences, including fractures [1]. Osteoporosis is defined as a systemic metabolic bone disorder, characterized by reduced bone strength and the deterioration of bone tissue [2]-[4]. Lots of older adults (aged above 50 years old) are undiagnosed with osteoporosis in its initial stages until the incidence of fractures, which is the most serious complication of osteoporosis [1]. Several common factors are associated with osteoporosis, such as hormone imbalance, the usage of specific medications, smoking, lack of physical activity, calcium and vitamin D deficiency, race, gastrointestinal diseases, rheumatoid arthritis, kidney diseases, and family history of osteoporosis [5]-[7].

Osteoporosis affects around 500 million older adults globally [8]. According to a World Health Organization (WHO) report, as a consequence of high disability and death rates caused by osteoporosis, high treatment expenses, and the low quality of patients' lives, it has emerged as the second most serious health concern after cardiovascular disorders worldwide [9]. The consequences of osteoporosis are commonly serious and frequently result in needing continuous medical care at the hospital [10]. It was expected that half of patients with osteoporosis are at high risk of different types of fractures [11]. Hip fracture is associated with an increased risk of mortality, with approximately 25% of patients dying within one year [12]. Osteoporotic fractures not only affect the quality of patients' lives but also have a major economic load on the health systems [13]. Without preventative procedures, osteoporosis-related fractures are likely to increase significantly. A recent study stated that the expected annual costs of osteoporosis and related fractures in the US are around $25 billion [8].

Therefore, further exploration is vital to discover a complete set of risk factors related to osteoporosis, as identifying the comorbidities of the condition in early stages before the presence of fractures and referring cases to specialized medical care provides more chances for a better diagnosis and treatment of the disease and saves medical expenses in surgeries and hospitalization [8]. However, the complexity of interactions between the comorbidities makes it challenging to accurately predict patients at high risk of osteoporosis using traditional statistical methods [14]. Furthermore, the rapid expansion of clinical datasets enables the study and prediction of disease patterns. Nevertheless, the representation of clinical datasets is complicated, leading to high-dimensional data. High-dimensional data increases the time and storage needed for performing experiments.

Recent advances in machine learning methods offer bright chances to improve risk prediction [14]. Machine learning techniques can handle high-dimensional

data and pick up on complex relationships between factors. This makes them useful for creating new ways to look at large amounts of data based on the idea of dimensionality reduction. Machine learning could also help us understand data better when it comes to multifactorial diseases. Machine learning has shown potential in several medical applications, including disorder diagnosis, treatment, and patient stratification [14].

In the present paper, we aim to develop a predictive model for predicting osteoporosis using a complete set of chronic disease data. We expect this model to serve as a reference for the implementation and integration of machine learning methods in osteoporosis health management. Our model combines unsupervised learning to find possible subtypes of patients and supervised learning to help predict clinical outcomes based on these subtypes. We have proposed a novel methodology that has achieved better results than classifiers alone by incorporating unsupervised learning—HDBSCAN clustering.

The rest of this paper follows this structure. In Section 2, related work will be discussed. Section 3 details the materials and methods, as well as the set of experiments undertaken. Section 4 presents the results, and the remainder of 5 discusses the implications of the findings. Finally, Section 6 concludes the paper and suggests directions for future research.

2. Related Work

Deploying various learning methods to uncover patterns and associations has garnered significant attention [15]. In this section, we briefly discuss the previous work that has exploited learning algorithms for the prediction of osteoporosis risk, which will help in assembling the proposed approach. In recent years, researchers have implemented several machine-learning approaches to improve osteoporosis prediction using clinical and demographic data. Several conventional machine learning methods, such as random forests, decision trees, and logistic regression, have shown reasonable achievement in classifying patients at high risk.

Khanna *et al.* [16] presented a decision support system that combined a machine learning framework and artificial intelligence for the prediction of osteoporosis. They presented a hybrid methodology that combines a feature selection approach with classification methods to enhance the identification of patients at high risk of osteoporosis. By deploying such a methodology, they achieved significant results. Furthermore, authors in [17] presented a predictive models using a multiple regression algorithm to predict bone mineral density tests using different features. They compared their results with the standard method - quantitative computed tomography (QCT) and their model demonstrated similar results to QCT. They then deployed a Logistic regression algorithm to predict osteoporosis. They achieved reasonable accuracy by using their proposed methodology. Moreover, Huang *et al.* [18] deployed a number of algorithms for feature selection and classification methods. They employed Mann-Whitney U test and the least absolute shrinkage and selection operator (LASSO) method to assess variables, then

deployed six machine learning algorithms for osteoporosis prediction. Their results indicated that the proposed methods enhanced the performance of machine learning techniques used to uncover the most suitable set of features.

Recently, various deep learning algorithms have been deployed to predict osteoporosis owing to their ability to construct accurate models from complex hidden interactions. Latest clinical research have built predictive models using algorithms such as Long Short-Term Memory Networks (LSTMs) and Deep Neural Networks. Such algorithms not only make accurate models but also they identify associations and relationships that might be missed using conventional machine-learning techniques.

Authors in [19] demonstrated the implementation of a novel DNN predictive model. Their model achieved an accuracy score of 75.4% and an area under the curve (AUC) of 0.848. They compared their model performance with several machine learning classifiers, Random Forest, Artificial Neural Network, K-Nearest Neighbor, and Support Vector Machine, in addition to a conventional regression model, named osteoporosis self-assessment tool (OST). Their model proved to be more efficient for timely diagnosis of osteoporosis in older people. Lin *et al.* [20] tackled the problem of osteoporosis diagnosis by developing predictive methods, *i.e.*, Artificial Neural Network, Random Forest, and Support Vector Machine and Logistic Regression. Genetic algorithm was used to adopt the suitable variables for predictive models. To assess the findings, they conducted Leave-One-Out Cross Validation. They stated that deploying Logistic Regression offers better results than other techniques used. A drawback of this study is that the size of the data is quite small; hence, the proposed model has to be tested using larger datasets.

The use of ensemble algorithms to improve osteoporosis prediction performance is also widespread. Methods such as boosting and stacking integrate several predictive models to decrease variance and bias. In [21], the authors used a mixed ensemble learning method that combined XGBoost and bagging to predict osteoporosis based on a number of risk factors, such as medical conditions and way of life. The results of the proposed methods achieved an accuracy of 88%, indicating their efficiency in classifying individuals at high risk. A major limitation of this work is that the sample size is quite small, which may restrict the robustness of the proposed methods across larger datasets. In the same way, authors in [14] suggested a number of different predictive models for finding osteoporosis early on using chronic disease datasets and variables related to those diseases. The Stacker model, which is made up of several classifiers such as logistic regression, AdaBoost classifier, and gradient boosting classifier, did the best of the predictive models.

Researchers have examined the use of unsupervised learning techniques to recognize potential patient subtypes, and supervised learning techniques to predict osteoporosis disease. Medical research has broadly deployed clustering techniques to group patients based on their characteristics. Although osteoporosis-based studies have not explored the use of such techniques, other clinical research fields

offer bright examples. The authors in [22] introduced a novel approach that combined consensus clustering with nearest-neighbor classification. They identified several distinct subtypes, which were then used to forecast various medical outcomes and enhance the accuracy of disease predictions. These novel methods demonstrated the potential for improving precision medicine by integrating unsupervised learning with supervised learning. In addition, authors [23] deployed machine learning methods to identify the likelihood of survival of heart failure patients. They proposed a combined approach that used supervised and unsupervised learning techniques. Unsupervised learning methods, namely Random Forest, XGBoost, and Decision Tree, were applied to group patients based on their clinical characteristics, where supervised learning methods were deployed for the prediction of the condition. Their results achieved high accuracy in prediction. Their findings emphasized the usefulness of using hybrid methods to enhance performance and predict heart failure.

These studies suggested that the integration of supervised learning and unsupervised learning algorithms could be valuable for predicting osteoporosis. This study extends the previous significant works by performing a hybrid methodology to explore the performance of a great number of unsupervised learning algorithms to find patient subtypes and supervised learning algorithms to predict the condition based on such subtypes. By leveraging the robustness of the methodology, the present work aims to enhance predictive accuracy and offer meaningful insights for personalized healthcare.

3. Methods and Materials

3.1. Study Population

The aim of this work was to apply a predictive model for osteoporosis risk among elderly patients. In this research, the dataset used is based on the German nationwide chronic disease data, which is all open-source and can be accessed from Dryad [24]. We obtained the data from 10,000 patients who had complete records. This cohort comprised an older population aged 65 years and above. These patients were associated with 10 different chronic diseases: hypertension, lipid metabolism, diabetes, coronary heart disease (CHD), cancer, chronic obstructive pulmonary disease (COPD), heart failure, stroke, chronic kidney disease, and osteoporosis. This data offers comprehensive information on attributes associated with osteoporosis development.

3.2. Data Preparation

The training and testing data were randomly split into 70% and 30% respectively. The proposed supervised learning methods will initially train the models using the training data to predict osteoporosis and afterward, they test the prediction based on the test data. There is always a class imbalance of data distribution which leads to a biasness towards the majority class. To address this concern, Synthetic Minority Oversampling Technique (SMOTE) was applied. This approach in-

volved replicating the minority class samples to balance the dataset, to reduce overfitting, training time, and enhance the proposed model accuracy.

3.3. Model Development and Validation

The proposed predictive model uses a hybrid method, combining unsupervised and supervised learning methods to improve osteoporosis prediction. We employed seven different machine learning methods to build predictive models. These methods are random forest, extra trees, gradient boosting, AdaBoost, XGBoost, LightGBM, and logistic regression. The initial step involves evaluating the use of these methods without implementing unsupervised learning to compare the results with our proposed model.

We developed the proposed model in two steps. In the first step, we performed principal component analysis (PCA) on the original data. We can find the most important principal components and, by extension, the number of dimensions in the data by looking at the PCA results in the form of scree plots. This transformation phase now represents each patient in a low-dimensional vector space, facilitating clustering and visualization. Patients were then clustered into subtypes. Then, we applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for identifying distinct subtypes. HDBSCAN is a complicated clustering algorithm that implemented to find subtypes of patients with varying densities and be more robust to parameter selection. Unlike other simple clustering algorithms, HDBSCAN allows for flexible cluster determination in complex datasets without the need for a prespecified number of clusters. HDBSCAN is a suitable clustering algorithm at finding distinct subtypes within various populations. This provides us with information about different subtypes of osteoporotic patients who might have different indicators and therefore, they might need different types of treatment.

The second step used such subtypes to construct predictive models. We deployed the same set of classification algorithms on each subtype for prediction. By training the model of each subtype independently, the proposed methodology capture different variables within each subtype, and therefore improve predictions.

This biphasic approach not only improves the performance accuracy of the predictive models, but also supports disease stratification and characterization, which helps doctors to dealing with patients based on their subtype. Integrating HDBSCAN with advanced classifiers emphasizes the potential of this methodology for clinical decision-making. A number of performance metrics were used to evaluate the performance of the proposed models, namely, accuracy, precision, recall, and F1-score.

3.4. Implementation and Computational Setup

This work was performed using Jupyter Notebook (Anaconda3) with Python (version 3.9) running on a MacBook Pro with macOS Big Sur (version 11.7.10),

equipped with 16 GB of RAM and an Intel® Core™ i9 CPU @ 2.3 GHz. The following Python packages were used:

- **Imbalanced-learn (version 0.12.3):** Used for implementing SMOTE technique.
- **XGBoost (version 2.1.2):** Implemented for the XGBoost classifier.
- **Lightgbm (version 4.5.0):** Used to implement the LightGBM classifier.
- **Scikit-learn (version 1.4.2):** Used for implementing other classifiers, PCA, evaluation metrics, and data splitting.

4. Results

This study utilized an open-source healthcare dataset from the Disease Analyzer database (IMS HEALTH) to assess the risk of osteoporosis among elderly patients in general practices, resulting in a cohort of 10,000 patients. 8707 (87%) of these patients have a diagnosis of osteoporosis, while 1293 (13%) do not. A total of 11 osteoporosis-relevant characteristics were considered. Table 1 outlines the characteristics distribution. To perform the experiments, the dataset was randomly partitioned: 70% (n = 7000) as the training set and 30% (n = 3000) as the testing set. We listed all evaluation metrics in percentage form, which included accuracy, precision, recall, and F1-score.

Table 1. Baseline characteristics of the dataset.

Characteristics	Overall (n = 10,000)	Osteoporotic (n = 8707)	Non-osteoporotic (n = 1293)
Male gender (%)	41.8	13.1	46.1
Age (mean)	76.9	79.5	76.5
Hypertension (%)	67.1	78.5	65.4
CHD (%)	25.7	33.8	24.5
Lipid_disorder (%)	41.2	51.0	39.7
Stroke (%)	6.3	8.3	6.0
Heart_failure (%)	15.7	25.2	14.3
Cancer (%)	17.1	22.3	16.4
Diabetes (%)	31.4	32.8	31.2
COPD (%)	12.9	20.6	11.7
Osteoporosis (%)	12.9	100.0	0.0
Chronic kidney disease (%)	13.0	17.2	12.4

Initially, we assessed and compared the performance of seven different classifiers without applying unsupervised learning algorithms. Table 2 illustrates the initial prediction performances of these algorithms. It is observed that the LightGBM classifier obtained the best performance, 82.3%, 82.3%, 82.3%, and

82.3% for accuracy, precision, recall, and F-score, respectively. On the other hand, logistic regression demonstrated the lowest accuracy score of 65%.

Table 2. Performance of using classification methods only.

Classifiers	Accuracy	Precision	Recall	F1-score
Random Forest	79%	76%	79%	77%
Extra Trees	80%	78%	80%	79%
Gradient Boosting	73%	71%	73%	69%
Ada Boost	68%	68%	68%	62%
XGBoost	81%	79%	81%	80%
LightGBM	82%	82%	82%	82%
Logistic Regression	65%	68%	65%	59%

To demonstrate the importance of osteoporosis subtyping, we combined unsupervised and supervised learning algorithms. Firstly, PCA was performed to facilitate the data visualization. Patients with similar diagnoses are placed together. Figure 1 shows the cumulative variance of the first ten principal components attained, as well as the data projected onto the first two components. After applying the transformation, each patient could now be thought of as being represented in a low-dimensional vector. This makes the clustering process faster and easier to understand. We then deployed the PCA matrix using the HDBSCAN clustering technique.

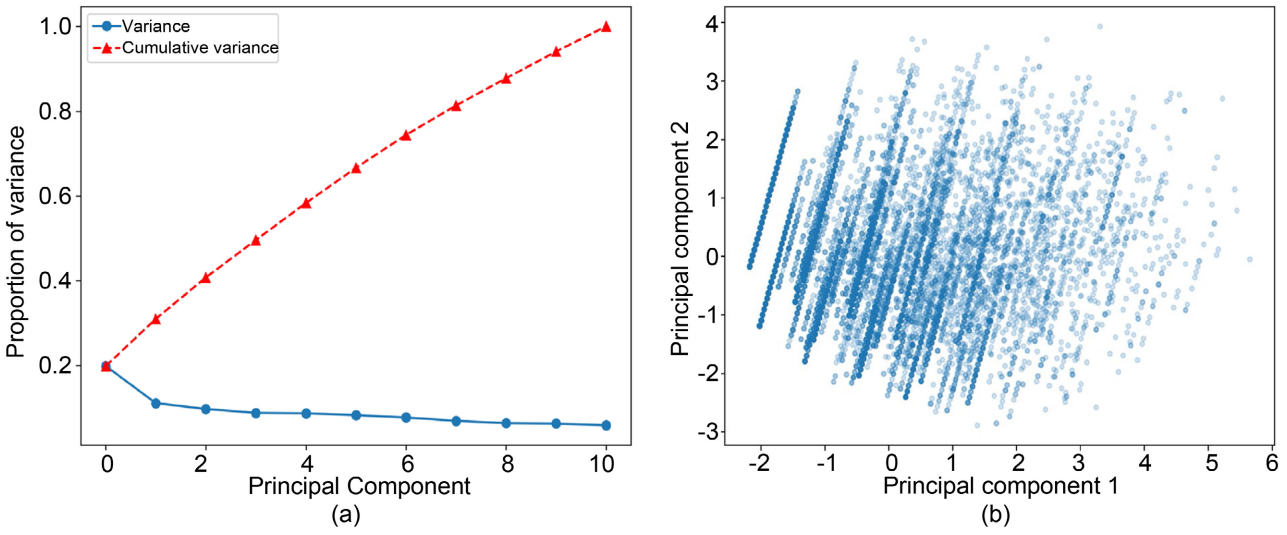


Figure 1. A low-dimensional representation of osteoporosis data using PCA. (a) The cumulative variance of principal components obtained. (b) The data projected onto the first two principal components.

The HDBSCAN was used for representation of same training set patients from the first experiment. The HDBSCAN parameters were examined and assessed by Silhouette coefficient measure. The HDBSCAN algorithm yielded seven distinct

subtypes. Clustering and distribution of patients in the subtypes are shown in Figure 2. Table 3 illustrates the descriptive statistics of the biomarkers for these seven subtypes. We identified that patients in subtype 1 were more associated with stroke (100%) and lipid disorder (88%). Patients in subtype 2 showed a higher association with COPD (100%). Heart failure (100%) and lipid disorder (49%) were the most common conditions associated with subtype 3. Patients in this subtype were mostly women (71%). Patients in subtype 4 did not show a significant association to any disease. Patients in subtype 5 were often diagnosed with cancer (100%). Patients in subtype 6 showed notable associations with chronic kidney disease (100%), lipid disorder (56%), and diabetes (54%). Patients in subtype 7 were enrich with various diseases related to chronic kidney disease (100%), heart failure (100%), lipid disorder (76%), CHD (73%) and diabetes (73%). All subtypes showed enrichment with hypertension ranging between 57% - 99%.

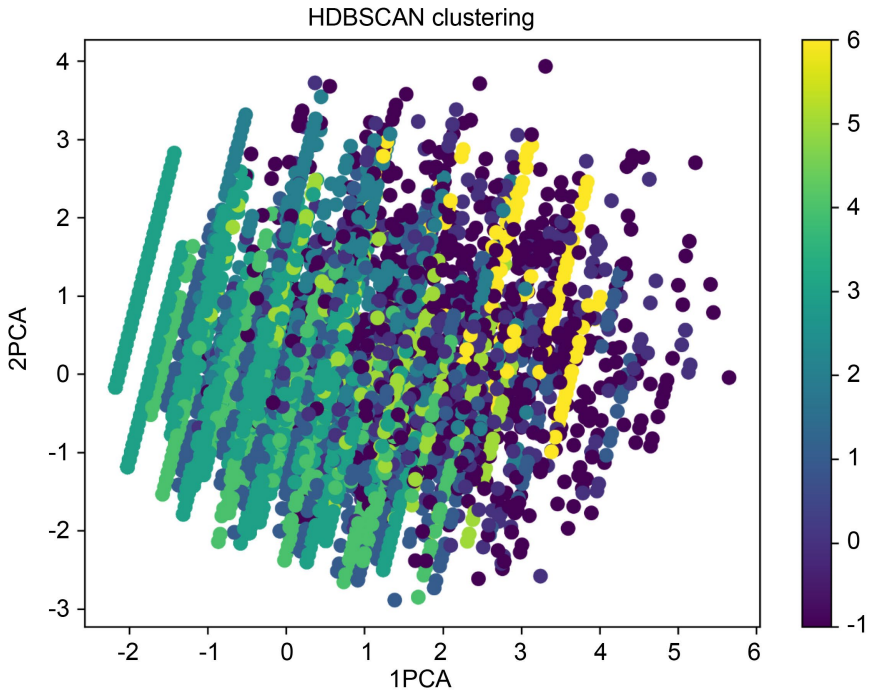


Figure 2. The HDBSCAN cluster analysis is based on the PCA matrix.

Following the preliminary assessment, we designated these subtypes for detailed refinement. We built predictive models for each subtype using the same set of classifiers deployed earlier. Then, we used core points to connect each point (patient) in the test set to its corresponding subtype and put it in the subtype of the core point that is closest to it by epsilon (eps). We identified 810 patients (8.1%) as outliers and did not assign them to any subtype. Table 4 displays the distribution of the training set and test set across subtypes. We used the predicative model of the assigned subtype to test the patient. Tables 5-11 illustrated the performance of the selected classification algorithms deployed to various subtypes.

Table 3. Subtype-wise distribution of characteristics.

Characteristics	Subtype 1 (n = 566)	Subtype 2 (n = 839)	Subtype 3 (n = 612)	Subtype 4 (n = 5596)	Subtype 5 (n = 929)	Subtype 6 (n = 464)	Subtype 7 (n = 184)
Male gender (%)	49.1	45.7	28.9	39.1	48.7	42.0	30.4
Age (maen)	78.9	75.7	80.5	75.9	76.4	77.4	80.0
Hypertension (%)	88.0	77.0	86.0	57.5	70.6	91.2	99.5
CHD (%)	40.5	33.7	41.5	15.7	23.0	27.4	72.8
Lipid disorder (%)	56.5	50.3	48.7	33.5	43.3	56.5	76.1
Stroke (%)	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Heart failure (%)	24.7	16.9	100.0	0.0	0.0	0.0	100.0
Cancer (%)	20.7	15.5	8.0	0.0	100.0	6.3	1.6
Diabetes (%)	43.3	32.9	33.3	23.9	29.5	53.9	72.8
COPD (%)	12.4	100.0	0.0	0.0	0.0	0.0	0.0
Osteoporosis (%)	15.6	19.4	19.6	9.6	14.2	12.5	22.8
Chronic kidney disease (%)	21.7	3.8	0.0	0.0	0.0	100.0	100.0

Table 4. Distribution of training and test sets across subtypes.

Subtypes	Number of training set (%)	Training set split	Number of test set (%)	Test set split
Subtype 1	400 (6.2%)	70.7%	166 (6%)	29.3%
Subtype 2	586 (9.1%)	69.8%	253 (9.2%)	30.2%
Subtype 3	420 (6.5%)	68.6%	192 (7%)	31.4%
Subtype 4	3889 (60.5%)	69.5%	1706 (61.8%)	30.5%
Subtype 5	677 (10.5%)	72.9%	252 (9.1%)	27.1%
Subtype 6	327 (5.1%)	70.5%	137 (5%)	29.5%
Subtype 7	130 (2%)	70.7%	54 (2%)	29.3%

Table 5. Performance of using clustering along with Random forest classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	88.0%	93.8%	88.0%	90.5%
Subtype 2	85.0%	91.5%	85.0%	88.1%
Subtype 3	82.8%	92.3%	82.8%	87.3%
Subtype 4	84.1%	88.1%	84.1%	86.0%
Subtype 5	86.1%	90.0%	86.1%	87.8%
Subtype 6	80.3%	84.2%	80.3%	82.1%
Subtype 7	87.0%	94.8%	87.0%	90.3%
Weighted average	84.4%	89.2%	84.4%	86.6%

Table 6. Performance of using clustering along with Extra tree classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	84.9%	88.2%	84.9%	86.5%
Subtype 2	85.4%	92.3%	85.4%	88.6%
Subtype 3	80.7%	85.6%	80.7%	83.0%
Subtype 4	83.7%	86.5%	83.7%	85.0%
Subtype 5	85.7%	89.3%	85.7%	87.3%
Subtype 6	78.8%	79.3%	78.8%	79.1%
Subtype 7	85.2%	87.6%	85.2%	86.3%
Weighted average	83.7%	87.0%	83.7%	85.2%

Table 7. Performance of using clustering along with Gradient Boosting classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	88.6%	98.2%	88.6%	92.9%
Subtype 2	88.9%	98.5%	88.9%	93.1%
Subtype 3	85.9%	97.3%	85.9%	91.1%
Subtype 4	87.9%	97.3%	87.9%	92.1%
Subtype 5	88.1%	93.6%	88.1%	90.4%
Subtype 6	85.4%	93.8%	85.4%	88.9%
Subtype 7	77.8%	77.8%	77.8%	77.8%
Weighted average	87.6%	96.6%	87.6%	91.6%

Table 8. Performance of using clustering along with AdaBoost classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	88.6%	100.0%	88.6%	93.9%
Subtype 2	88.9%	97.4%	88.9%	92.6%
Subtype 3	85.9%	98.9%	85.9%	92.0%
Subtype 4	87.9%	97.3%	87.9%	92.1%
Subtype 5	87.7%	96.8%	87.7%	91.8%
Subtype 6	85.4%	97.8%	85.4%	90.9%
Subtype 7	79.6%	84.7%	79.6%	82.1%
Weighted average	87.6%	97.3%	87.6%	92.0%

Noteworthy observations show that the use of HDBSCAN clustering had the best results with the learning classifiers in all cases compared to using classifiers only. The results showed that logistic regression was the best classifier. It got average weighted scores of 87.8% for accuracy, 98.6% for precision, 87.8% for recall, and 92.7% for F1-score. This model had a high precision and F1-score, which

meant it could correctly classify positive cases and find a balance between precision and recall. Gradient Boosting and AdaBoost, based on subtypes classification, also had superior performances, with accuracies of 87.6%. As shown in Figure 3, we compare the results of the models using classification methods only and deploying clustering along with the classifiers.

Table 9. Performance of using clustering along with XGBoost classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	86.7%	90.3%	86.7%	88.3%
Subtype 2	83.4%	86.7%	83.4%	84.9%
Subtype 3	83.9%	91.7%	83.9%	87.4%
Subtype 4	86.3%	92.1%	86.3%	88.9%
Subtype 5	84.1%	87.2%	84.1%	85.5%
Subtype 6	83.9%	86.7%	83.9%	85.2%
Subtype 7	83.3%	82.5%	83.3%	82.9%
Weighted average	85.5%	90.5%	85.5%	87.8%

Table 10. Performance of using clustering along with LightGBM classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	88.6%	93.6%	88.6%	90.7%
Subtype 2	85.4%	90.4%	85.4%	87.7%
Subtype 3	84.9%	95.1%	84.9%	89.6%
Subtype 4	87.5%	94.8%	87.5%	90.7%
Subtype 5	84.9%	88.6%	84.9%	86.6%
Subtype 6	83.2%	89.6%	83.2%	86.0%
Subtype 7	81.5%	79.8%	81.5%	80.5%
Weighted average	86.6%	93.2%	86.6%	89.5%

Table 11. Performance of using clustering along with Logistic Regression classifier.

Subtypes	Accuracy	Precision	Recall	F1-score
Subtype 1	88.6%	100.0%	88.6%	93.9%
Subtype 2	88.9%	99.6%	88.9%	93.8%
Subtype 3	86.5%	100.0%	86.5%	92.7%
Subtype 4	88.1%	98.2%	88.1%	92.6%
Subtype 5	87.3%	98.3%	87.3%	92.5%
Subtype 6	85.4%	100.0%	85.4%	92.1%
Subtype 7	85.2%	96.1%	85.2%	90.3%
Weighted average	87.8%	98.6%	87.8%	92.7%

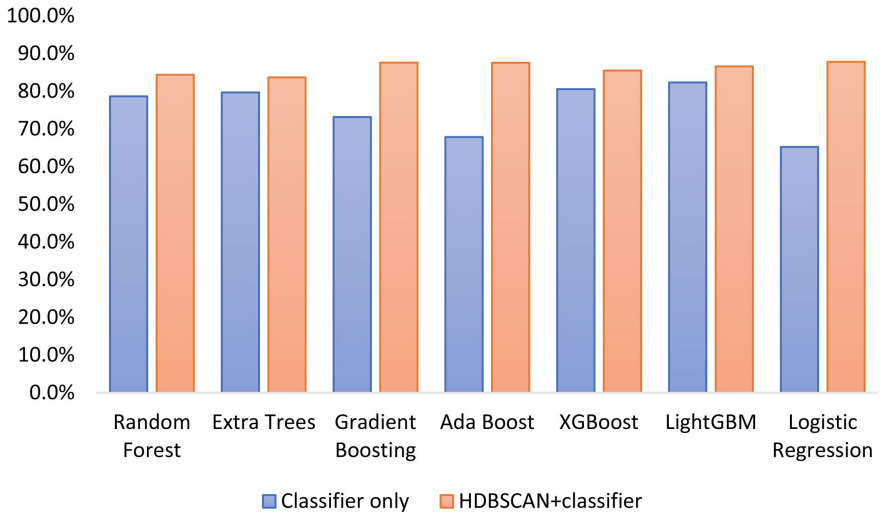


Figure 3. Evaluation of the classification accuracy for all predictive models.

5. Discussion

Osteoporosis is a public condition marked by decreased bone density, which likely increases fracture risk. While indicators usually remain unobserved until a fracture occurs, the need for early diagnosis and prevention becomes essential. Adverse outcomes, increased mortality, and high healthcare expenses are associated with osteoporosis-related fractures, particularly hip fractures. This renders predicting osteoporosis vital, not only for ensuring better medical management but also for dealing with public health and economic challenges.

As a result, predicting osteoporosis early can pave the way for timely interventions that can help slow or even reverse bone loss. This approach effectively lowers the risk of various fractures, enhances the quality of life for older adults, promotes greater independence, and lowers healthcare costs. Strategies, including lifestyle adjustments and clinical treatments are considerably more effective when osteoporosis is recognized in its initial phases.

There is a lot of interest in applying machine learning and data mining techniques for discovering patterns and correlations [15]. Numerous studies have explored the prediction of osteoporosis, utilizing a variety of machine learning techniques. Most of the existing studies rely mainly on bone density scans. However, these assessments only take place when symptoms manifest, potentially delaying early intervention.

In various real-world situations, obtaining detailed clinical datasets can be challenging. Therefore, the first step in developing an effective predictive model for osteoporosis is to select an appropriate and up-to-date dataset. The dataset should include both positive (osteoporosis) and negative (non-osteoporosis) cases to reflect the types of cases the model will encounter in real-world applications. In this research, we used the nationwide chronic disease dataset from Germany, which is recent, well-structured, and free from extreme noise, making it appropriate for reliable analysis.

The findings of this paper emphasize the potential of combining unsupervised and supervised machine learning techniques to enhance osteoporosis prediction and offer useful characterizations of patients. This new method builds on earlier work by showing how important it is to use clustering methods, like HDBSCAN, to find unique subtypes in a large group of different patients and then use those subtypes to make the models better at making predictions. We investigated the whole elderly population using classifiers only; again, we applied the same classifiers to osteoporosis clusters to see how predictive models might change at the subtype level.

The results show the effectiveness of combining dimensionality reduction and clustering techniques with traditional classifiers in clinical data analysis. Deploying PCA as a preprocessing phase not only simplified data visualization and reduced dimensionality but also enhanced clustering efficiency. HDBSCAN was used to find seven completely separate groups of patients, each with its own set of variables that were linked to it. This stratification exposed patterns that may otherwise remain hidden in medical datasets. For instance, Subtype 1 patients showed significant associations with stroke and lipid disorders, while Subtype 6 was considerably enriched with chronic kidney disease, lipid disorders, and diabetes. Prior studies that link conditions like heart and metabolic diseases to osteoporosis support these findings [16] [18]. The enrichment of hypertension in all subtypes further confirms its role as an independent risk factor, as stated in the existing literature [25].

According to the research, using subtypes analysis along with classifiers makes predictions much more accurate than using classifiers alone. Initially, the logistic regression classifier did not do well with the dataset without prior subtyping (only 65% accurate), but it did very well when subtype-specific model training was added, with a weighted average accuracy of 87.8%. This emphasizes the importance of classifying patients into clusters to capture the differences in disease progression and risk factors. These findings are consistent with prior research indicating the advantages of deploying a hybrid approach when analyzing complex medical datasets [22].

By adding unsupervised learning methods to classification methods, our approach makes them better, leading to better accuracy, precision, and recall performance metrics. The performance accuracies of the subtypes range between 85.2% and 88.9%, highlighting that the variation in their performance levels is relatively minor. This narrow range suggests that all subtypes perform at a comparable level, with no single subtype showing a significantly higher accuracy than the others. Using HDBSCAN clustering, which doesn't require deciding ahead of time on the number of clusters, also makes the approach more flexible and adaptable, getting around problems that were seen in studies that used traditional clustering methods [22] [23]. The implications of the findings presented in this study are significant for personalized medicine and health management. Patient subtypes identification aids stakeholders for timely interventions and resource allocations.

6. Conclusions

This study represents an analysis of a cohort of patients diagnosed with osteoporosis. It contributes towards a better understanding of the occurrence of osteoporosis as a health condition, with assessment of factors that can lead to osteoporosis, and the associated effects impacting the quality of patients' lives. This research evaluated the performance of subtype analysis with various classification methods to enhance the prediction of osteoporosis. This study introduced a novel hybrid machine learning approach to stratify patients with osteoporosis into subtypes, leading to improved osteoporosis detection. The proposed methodology enhanced prediction accuracy compared to using classifiers only by using HDBSCAN to cluster patients together into distinct subtypes. It then deployed different classifiers on such subtypes. Although logistic regression didn't perform properly on the dataset without clustering at first, but it achieved much better results when applied to subtypes. These findings highlight the potential of leveraging patient subtyping to improve predictive models' performance accuracy and facilitate personalized healthcare plans for osteoporosis.

While this study attained positive results, it has numerous limitations. Although HDBSCAN has demonstrated effectiveness in identifying meaningful subtypes, outlier management (8.1% of the dataset) remains a persistent difficulty. Future investigations could explore alternative clustering approaches to better deal with outliers and further improve subtype classification. Furthermore, the study depended upon chronic conditions, which may not capture all relevant osteoporosis-related risk factors, such as dietary habits or physical activity levels.

In conclusion, this study used a biphasic methodology to characterize and predict patients with osteoporosis at the subtype level. By addressing existing limitations and expanding its scope, this methodology holds promise for enhancing precision medicine and advancing medical outcomes in osteoporosis management.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Kemmak, A.R., Rezapour, A., Jahangiri, R., Nikjoo, S., Farabi, H. and Soleimanpour, S. (2020) Economic Burden of Osteoporosis in the World: A Systematic Review. *Medical Journal of the Islamic Republic of Iran*, **34**, 154.

[2] Aydin Ozturk, P., Arac, E., Ozturk, U. and Arac, S. (2021) Estimation of Bone Mineral Density with Hounsfield Unit Measurement. *British Journal of Neurosurgery*, **38**, 464-467. https://doi.org/10.1080/02688697.2021.1888877

[3] Lu, B. and Zhang, L. (2023) Association of a History of Gestational Diabetes Mellitus with Osteoporosis, Bone Mineral Density, and Trabecular Bone Score in Postmenopausal Women. *Diabetology & Metabolic Syndrome*, **15**, Article No. 215. https://doi.org/10.1186/s13098-023-01194-8

[4] Wu, Y., Chao, J., Bao, M. and Zhang, N. (2023) Predictive Value of Machine Learning on Fracture Risk in Osteoporosis: A Systematic Review and Meta-Analysis. *BMJ*

Open, **13**, e071430. https://doi.org/10.1136/bmjopen-2022-071430

[5] Kong, S.H. (2024) Sex/Gender Differences in Osteoporosis. In: *Sex/Gender-Specific Medicine in Clinical Areas*, Springer, 277-286. https://doi.org/10.1007/978-981-97-0130-8_13

[6] Xu, P., Ge, J., Jiang, H., Lin, Y., Ye, Y., Huang, X., *et al.* (2023) Gastrointestinal Disease Is an Important Influencing Factor of Osteoporosis Fracture: A Retrospective Study in Chinese Postmenopausal Women. *BMC Musculoskeletal Disorders*, **24**, Article No. 659. https://doi.org/10.1186/s12891-023-06765-4

[7] Rahim, F., Zaki Zadeh, A., Javanmardi, P., Emmanuel Komolafe, T., Khalafi, M., Arjomandi, A., *et al.* (2023) Machine Learning Algorithms for Diagnosis of Hip Bone Osteoporosis: A Systematic Review and Meta-Analysis Study. *BioMedical Engineering Online*, **22**, Article No. 68. https://doi.org/10.1186/s12938-023-01132-9

[8] Albuquerque, G.A., Carvalho, D.D.A., Cruz, A.S., Santos, J.P.Q., Machado, G.M., Gendriz, I.S., *et al.* (2023) Osteoporosis Screening Using Machine Learning and Electromagnetic Waves. *Scientific Reports*, **13**, Article No. 12865. https://doi.org/10.1038/s41598-023-40104-w

[9] Xiao, P.-L., Cui, A.-Y., Hsu, C.-J., Peng, R., Jiang, N., *et al.* (2022) Global, Regional Prevalence, and Risk Factors of Osteoporosis According to the World Health Organization Diagnostic Criteria: A Systematic Review and Meta-Analysis. *Osteoporosis International*, **33**, 2137-2153. https://doi.org/10.1007/s00198-022-06454-3

[10] Ramchand, S.K. and Leder, B.Z. (2023) Sequential Therapy for the Long-Term Treatment of Postmenopausal Osteoporosis. *The Journal of Clinical Endocrinology & Metabolism*, **109**, 303-311. https://doi.org/10.1210/clinem/dgad496

[11] Al-Moaibed, G., Al Hamam, N., Alfayez, E., Alfayez, E., Al-Mubaddil, M. and Alramadhan, N. (2020) Prevalence and Risk Factors for Osteoporotic Fracture among Adults with Comorbidities in Al-Ahsaa, Saudi Arabia. *Journal of Family Medicine and Primary Care*, **9**, 877-882. https://doi.org/10.4103/jfmpc.jfmpc_982_19

[12] Sing, C.W., Lin, T.C., Bartholomew, S., Bell, J.S., Bennett, C., Beyene, K., Bosco-Levy, P., Bradbury, B.D., Chan, A.H.Y., Chandran, M. and Cooper, C. (2023) Global Epidemiology of Hip Fractures: Secular Trends in Incidence Rate, Post-Fracture Treatment, and All-Cause Mortality. *Journal of Bone and Mineral Research*, **38**, 1064-1075.

[13] Moayyeri, A., Warden, J., Han, S., Suh, H.S., Pinedo-Villanueva, R., Harvey, N.C., *et al.* (2023) Estimating the Economic Burden of Osteoporotic Fractures in a Multinational Study: A Real-World Data Perspective. *Osteoporosis International*, **34**, 2121-2132. https://doi.org/10.1007/s00198-023-06895-4

[14] Tu, J., Liao, W., Liu, W. and Gao, X. (2024) Using Machine Learning Techniques to Predict the Risk of Osteoporosis Based on Nationwide Chronic Disease Data. *Scientific Reports*, **14**, Article No. 5245. https://doi.org/10.1038/s41598-024-56114-1

[15] Garg, U., Shukla, A.K., Singh, H., Sharma, N. and Kaur, E.S. (2020) Use of Machine Learning in the Pattern Finding. *International Journal of Recent Technology and Engineering*, **9**, 527-531. https://doi.org/10.35940/ijrte.a1237.059120

[16] Khanna, V.V., Chadaga, K., Sampathila, N., Chadaga, R., Prabhu, S., K S, S., *et al.* (2023) A Decision Support System for Osteoporosis Risk Prediction Using Machine Learning and Explainable Artificial Intelligence. *Heliyon*, **9**, e22456. https://doi.org/10.1016/j.heliyon.2023.e22456

[17] Nam, K.H., Seo, I., Kim, D.H., Lee, J.I., Choi, B.K. and Han, I.H. (2019) Machine Learning Model to Predict Osteoporotic Spine with Hounsfield Units on Lumbar

Computed Tomography. *Journal of Korean Neurosurgical Society*, **62**, 442-449. https://doi.org/10.3340/jkns.2018.0178

[18] Huang, C., Hu, J., Tan, K., Zhang, W., Xu, T. and Yang, L. (2022) Application of Machine Learning Model to Predict Osteoporosis Based on Abdominal Computed Tomography Images of the Psoas Muscle: A Retrospective Study. *BMC Geriatrics*, **22**, Article No. 796. https://doi.org/10.1186/s12877-022-03502-9

[19] Qiu, C., Su, K., Luo, Z., Tian, Q., Zhao, L., Wu, L., *et al*. (2024) Developing and Comparing Deep Learning and Machine Learning Algorithms for Osteoporosis Risk Prediction. *Frontiers in Artificial Intelligence*, **7**, Article 1355287. https://doi.org/10.3389/frai.2024.1355287

[20] Lin, Y., Chu, C., Hung, K., Lu, C., Bednarczyk, E.M. and Chen, H. (2022) Can Machine Learning Predict Pharmacotherapy Outcomes? An Application Study in Osteoporosis. *Computer Methods and Programs in Biomedicine*, **225**, Article 107028. https://doi.org/10.1016/j.cmpb.2022.107028

[21] Irmawati, I., Herdit Juningsih, E. and Yanto, Y. (2024) Predictive Modeling of Osteoporosis Risk Factors Using XGBoost and Bagging Ensemble Technique. *Journal Medical Informatics Technology*, **2**, 6-10. https://doi.org/10.37034/medinftech.v2i1.27

[22] Alyousef, A.A., Nihtyanova, S., Denton, C., Bosoni, P., Bellazzi, R. and Tucker, A. (2018) Nearest Consensus Clustering Classification to Identify Subclasses and Predict Disease. *Journal of Healthcare Informatics Research*, **2**, 402-422. https://doi.org/10.1007/s41666-018-0029-6

[23] Zaman, S.M.M., Qureshi, W.M., Raihan, M.M.S., Shams, A.B. and Sultana, S. (2021) Survival Prediction of Heart Failure Patients Using Stacked Ensemble Machine Learning Algorithm. 2021 *IEEE International Women in Engineering* (*WIE*) *Conference on Electrical and Computer Engineering* (*WIECON-ECE*), Dhaka, 4-5 December 2021, 117-120. https://doi.org/10.1109/wiecon-ece54711.2021.9829577

[24] Karel, K. (2016) Data from: Prevalence of Chronic Diseases among Older Patients in German General Practices. https://datadryad.org/stash/dataset/doi:10.5061/dryad.qh0h1

[25] Huang, Y. and Ye, J. (2024) Association between Hypertension and Osteoporosis: A Population-Based Cross-Sectional Study. *BMC Musculoskeletal Disorders*, **25**, Article No. 434. https://doi.org/10.1186/s12891-024-07553-4