JOURNAL OF DECISION MAKING AND HEALTHCARE
Volume 2 (2025), No. 3, 210–223
https://doi.org/10.69829/jdmh-025-0203-ta05
Tulipa Opera Scholarum

MACHINE LEARNING-BASED PREDICTION AND RISK FACTOR ANALYSIS OF OSTEOPOROSIS

WARAGUNT WARATAMRONGPATAI[1], KRITTIN NARAVEJSAKUL[1], WATCHARAPORN CHOLAMJIAK[2], AND PHATCHARAPON UDOMLUCK[1,*]

[1] *School of Medicine, University of Phayao, Phayao 56000, Thailand*
[2] *Department of Mathematics, School of Science, University of Phayao, Phayao 56000, Thailand*

Abstract. Background/Objectives: Osteoporosis is a silent and progressive skeletal disorder characterized by reduced bone mineral density and an increased risk of fractures. Early identification of high-risk individuals is essential to enable timely interventions and prevent complications. This study aimed to develop and evaluate a machine learning–based framework for osteoporosis risk prediction by integrating demographic, medical history, and lifestyle features, with particular focus on feature reduction strategies to balance accuracy, interpretability, and efficiency. Methods: A dataset comprising 18 clinically relevant predictors was analyzed. Predictors were systematically reduced through feature selection, resulting in four feature sets (Case1–4). Five supervised learning models were implemented in MATLAB R2023b, including Fine Tree, Opti-mizable Support Vector Machine (SVM), Efficient Logistic Regression, Neural Network, and Optimizable Naïve Bayes. Model performance was assessed using accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC). Results: The Fine Tree model under Case4, developed using only two predictors (Age and Medications), achieved the best balance between accuracy and computational efficiency, with 91.4% accuracy, 82.8% sensitivity, 100% specificity, an F1-score of 90.6%, and an AUC of 0.9208. The Optimizable SVM (Case3/4) achieved the highest predictive capability, with accuracy ranging from 90.9-91.4% and an F1-score up to 90.6%, but required substantially greater computational cost. Neural Networks and Optimizable Naïve Bayes demonstrated competitive AUC values but were constrained by either sensitivity or longer training times. Conclusions: Machine learning models with minimal feature sets can achieve clinically meaningful performance in osteoporosis risk prediction. The findings suggest that streamlined models such as Fine Tree (Case4) are well suited for rapid and resource-efficient screening, supporting their potential integration into re-al-world healthcare settings for early risk assessment and preventive care.

Keywords. Osteoporosis, Machine Learning, Risk Factor Analysis, Predictive Modeling, Clinical Decision Support.

1. Introduction

Osteoporosis is a chronic skeletal disorder characterized by reduced bone mineral density (BMD) and deterioration of bone microarchitecture, leading to increased bone fragility and fracture risk [1]. Often termed a "silent disease," it progresses asymptomatically until a fracture occurs, most commonly at the hip, spine, or wrist [2]. These fractures are associated with chronic pain, disability, diminished quality of life, and increased mortality [3]. Globally, osteoporosis constitutes a major public health challenge, with prevalence and healthcare burden expected to rise substantially in aging populations [4, 5].

In Thailand, the burden of osteoporosis among older adults, particularly women, is especially pronounced. The prevalence at the lumbar spine nearly doubles from 22.6% to 39.4% and at the femoral neck from 10.3% to 20.1% between ages 55-59 and 60-64 [6]. This sharp increase among older women

*Corresponding author.
E-mail address: waragunt.wa@up.ac.th (W. Waratamrongpatai), krittin.na@up.ac.th (K. Naravejsakul), watcharaporn.ch@up.ac.th (W. Cholamjiak), phatcharapon.ud@up.ac.th (P. Udomluck)
Accepted: November 06, 2025.

highlights the urgency of implementing targeted screening and preventive strategies before the age of 60. Bone mineral density (BMD) measurement remains the diagnostic gold standard [7], while fracture risk assessment tools such as FRAX® are widely employed to improve early detection [8, 9]. However, FRAX demonstrates only moderate sensitivity and omits several relevant risk factors, such as vitamin D deficiency [10], underscoring the necessity for more comprehensive predictive approaches tailored to older adults.

A central challenge in managing osteoporosis lies in its asymptomatic progression and the limited accessibility of dual-energy X-ray absorptiometry (DXA) in many regions [11], resulting in underdiagnosis until advanced stages. Risk factor–based predictive modeling therefore represents a cost-effective, non-invasive alternative to identify high-risk individuals [12]. Both non-modifiable (age, sex, ethnicity, family history) [13] and modifiable factors (low calcium and vitamin D intake, smoking, alcohol use, physical inactivity, and chronic diseases) [14, 15] influence susceptibility. Early assessment of these risk factors enables timely preventive strategies—including dietary optimization, weight-bearing exercise, pharmacological treatments, and patient education [5, 11]—which significantly reduce fracture incidence, improve out-comes, and lower healthcare costs. Furthermore, risk identification supports shared decision-making [16], empowers patients to adopt healthier lifestyles [9], and facilitates timely initiation of pharmacological therapy in high-risk populations [17], ultimately reducing fracture burden [18] and mitigating long-term economic costs [19].

Recent advances in data science and artificial intelligence provide new opportunities to strengthen osteoporosis prevention. Machine learning (ML) models can capture complex interactions among demographic, clinical, and lifestyle variables, yielding accurate risk predictions [20, 21]. Incorporating such approaches into clinical practice holds promise for improving screening efficiency, supporting decision-making, and enabling early, targeted interventions to lessen the global impact of osteoporosis.

2. Materials and Methods

2.1. **Dataset description.** The dataset used in this study was obtained from a publicly accessible source on the Kaggle platform, titled "Lifestyle Factors Influencing Osteoporosis" [22]. It includes information from 1,958 individuals and is designed to explore the relationship between life-style behaviors, demographic profiles, and clinical conditions in relation to osteoporosis. A total of 18 features were extracted and categorized into four groups: demographic variables (age, sex, weight, and height), lifestyle behaviors (smoking, alcohol consumption, and physical activity), medical history (including diabetes, rheumatoid arthritis, and pro-longed steroid use), and encoded categorical variables such as race/ethnicity and comorbidities. All categorical variables were converted into binary format using one-hot encoding. The outcome variable is binary, indicating the presence or absence of osteoporosis, and was used as the target for classification.

This dataset provides not only a reliable and realistic basis for evaluating machine learning models in the context of osteoporosis risk prediction, but also offers practical insights that can serve as a guideline for structuring real-world data collection in clinical and community healthcare settings. All procedures and analyses in this study were re-viewed and approved by the Human Research Ethics Committee of the University of Phayao, Thailand.

2.2. **Preprocessing.** Prior to model development, data preprocessing steps were applied to ensure completeness and suitability for analysis. Missing values were handled using a combination of median imputation for numerical variables and mode imputation for categorical variables to preserve distributional characteristics. Categorical features were transformed into numerical format through one-hot encoding [23]. A correlation matrix was computed among all features and the target variable to examine potential multicollinearity and identify relationships between predictors [24]. This exploratory step

provided insights into underlying associations and guided subsequent modeling decisions. To evaluate model performance, the dataset was partitioned into training and test subsets using a 90:10 split ratio, with stratification to maintain the original class distribution of the outcome variable [25].

2.3. **Feature selection.** To improve model interpretability and mitigate the risk of overfitting, four complementary feature selection approaches were employed: Logistic Regression, LASSO regularization, Random Forest, and correlation analysis. Logistic Regression was used to identify statistically significant predictors through odds ratios and p-values, highlighting variables with strong clinical relevance [26]. LASSO regularization [27] provided additional refinement by penalizing less informative coefficients and promoting sparsity in the feature space. Random Forest, as a tree-based ensemble method, was applied to capture nonlinear relationships and rank feature importance based on permutation error reduction [28]. Finally, correlation analysis was conducted to detect potential multicollinearity and to ensure that selected predictors contributed unique information [24]. By integrating results across these four methods, the final set of features was determined with greater robustness, balancing both statistical significance and predictive utility for osteoporosis risk modeling.

2.4. **Model development.** For the development of predictive models, five supervised machine learning classifiers were implemented using MATLAB R2023b to evaluate their effectiveness in osteoporosis risk prediction. These included Fine Tree [29] Optimizable Support Vector Machine (SVM) [30], Efficient Logistic Regression [26], Neural Network [31], and Optimizable Naïve Bayes [32]. Each model was selected based on its established suitability for classification tasks, as well as its ability to capture both linear and nonlinear relationships within the data. This diverse set of algorithms provided a balanced framework for benchmarking predictive performance, interpretability, and computational efficiency in the context of clinical decision support.

2.5. **Evaluation metrics.** Model performance was evaluated using multiple classification metrics, including accuracy, sensitivity, specificity, F1-score, and the area under the receiver operating characteristic curve (AUC) [33, 34]. These measures provide a comprehensive assessment of predictive capability and clinical relevance, particularly in identifying high-risk individuals for osteoporosis. The metrics are defined as follows:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

$$\text{Sensitivity}(\text{Recall}, \text{TPR}) = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

$$\text{Specificity}(\text{TNR}) = \frac{\text{TN}}{(\text{TN} + \text{FP})}$$

$$\text{F1} - \text{score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

where $\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$ and $\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$.
Definition of terms:
TP (True Positive): correctly predicted positive cases (e.g., individuals correctly classified as high risk).
TN (True Negative): correctly predicted negative cases (e.g., individuals correctly classified as low risk).
FP (False Positive): negative cases incorrectly predicted as positive (e.g., low-risk individuals misclassified as high risk).
FN (False Negative): positive cases incorrectly predicted as negative (e.g., high-risk individuals misclassified as low risk).

3. Results

3.1. **Correlation analysis.** A correlation analysis was performed among all input features and the target variable (osteoporosis status). The Pearson correlation matrix indicated that age had the strongest positive correlation with osteoporosis ($r = 0.69$). In contrast, variables such as physical activity, body weight, and alcohol consumption exhibited weak or negative correlations with the target. Some moderate negative inter-correlations were also observed among encoded race/ethnicity and medical condition variables. These findings provide preliminary insights into linear relationships among features and their potential relevance to osteoporosis prediction. A summary of pairwise correlation coefficients is presented in Figure 1.

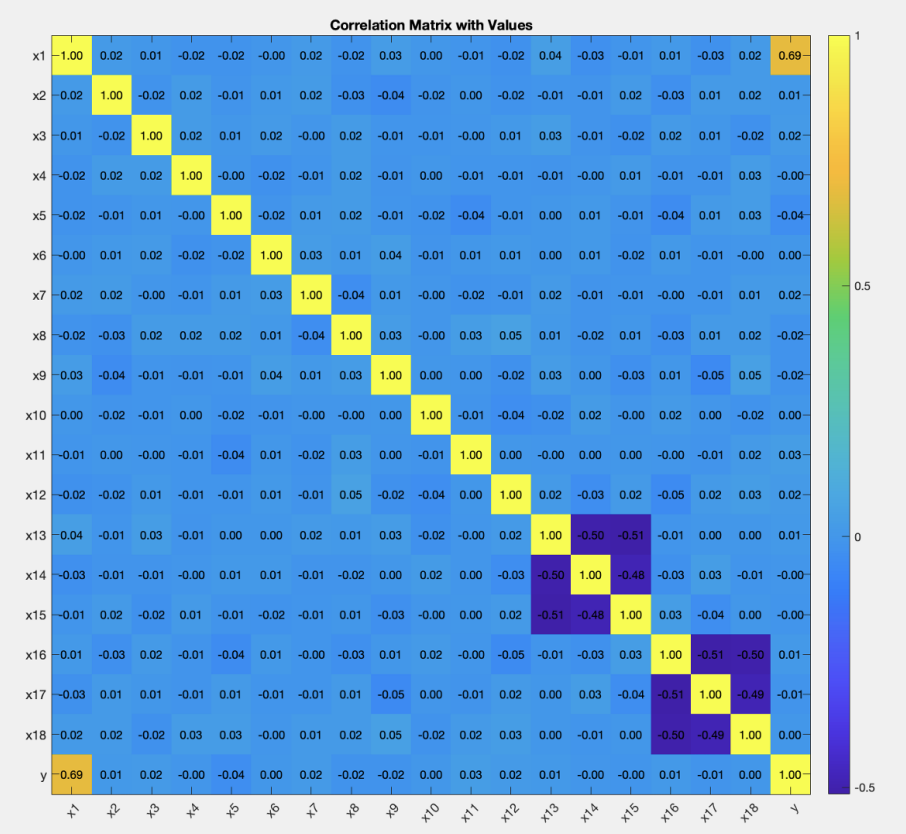


Figure 1. Pearson correlation matrix of all predictors and the target variable (Osteoporosis)

For clarity, the variables used in the correlation analysis are defined as follows:

x1 = Age
x2 = FamilyHistory
x3 = BodyWeight
x4 = CalciumIntake
x5 = VitaminDIntake
x6 = Smoking
x7 = Medications
x8 = PriorFractures
x9 = Race Ethnicity Asian
x10 = MedicalConditions RheumatoidArthritis

x11 = Gender
x12 = HormonalChanges
x13 = PhysicalActivity
x14 = AlcoholConsumption
x15 = Race Ethnicity AfricanAmerican
x16 = Race Ethnicity Caucasian
x17 = MedicalConditions Hyperthyroidism
x18 = MedicalConditions None
y = Osteoporosis (target variable)
These variable definitions correspond to the axes shown in the correlation matrix (Fig. 1) and are used consistently throughout the analysis.

3.2. **Feature importance.** To identify the most relevant predictors, logistic regression analysis was performed to evaluate the contribution of individual features to the outcome. The regression coefficients, adjusted odds ratios (AdjOR), and corresponding 95% confidence intervals were estimated to determine statistical significance and the strength of association. Features with significant coefficients were considered as strong predictors for model construction. The results are summarized in Table 1, while the overall classification performance of the proposed model is illustrated in Figs. 2 (a,b), showing both the confusion matrix and ROC curve.

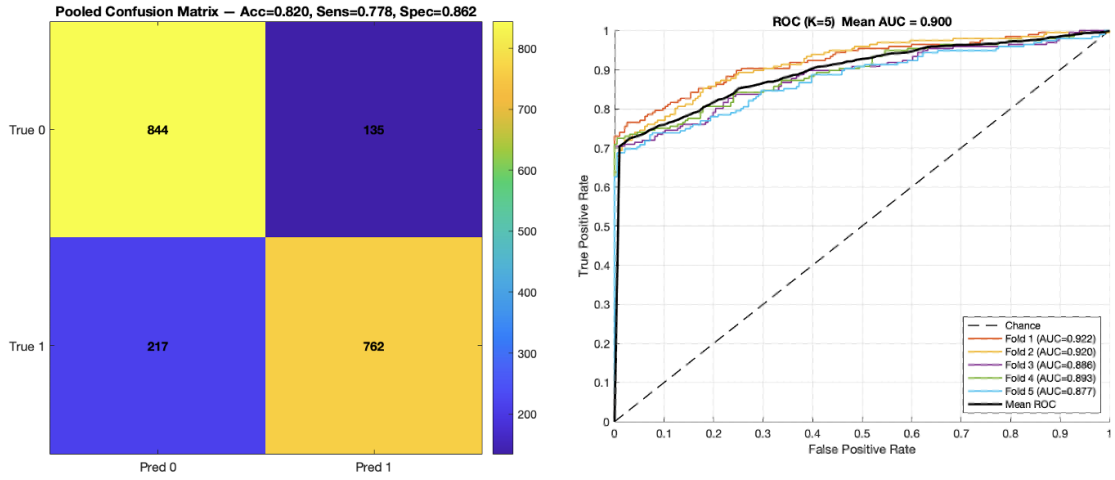


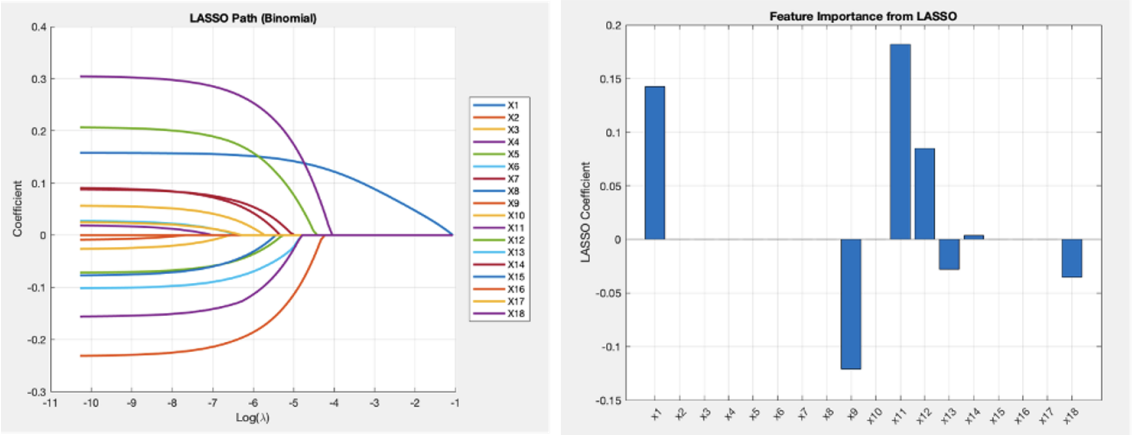
FIGURE 2. Model performance evaluation of logistic regression. (a) shows the pooled confusion matrix with overall accuracy of 0.820, sensitivity of 0.778, and specificity of 0.862. (b) presents the ROC curve with 5-fold cross-validation, achieving a mean AUC of 0.877

Note: SE = standard error; AdjOR = adjusted odds ratio; CI = confidence interval; Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$. Model summary: 1,958 observations, 1,941 degrees of freedom; Dispersion = 1; χ^2-statistic vs. constant model = 1.34×10^3, $p < 0.001$. Variable x18 (MedicalConditions None) was excluded from the logistic regression model as the reference category in the one-hot encoding scheme to avoid multicollinearity. Consequently, all other medical condition variables were interpreted relative to this baseline group.

LASSO regularization, random forest algorithm were employed to identify the most influential predictors of osteoporosis. These methods were selected for their complementary strengths in feature

TABLE 1. Estimated coefficients and adjusted odds ratios from logistic regression analysis

Variable	Beta (Estimate)	SE	t-value	p-value	AdjOR	95% CI (Lower–Upper)	Significance
x1	3.373	0.165	20.402	1.61×10^{-92}	29.16	21.09-40.33	$***$
x2	0.010	0.068	0.143	0.886	1.010	0.884-1.154	
x3	-0.036	0.068	-0.532	0.595	0.964	0.844-1.102	
x4	0.014	0.068	0.205	0.838	1.014	0.887-1.159	
x5	0.045	0.068	0.668	0.505	1.046	0.916-1.196	
x6	-0.116	0.068	-1.703	0.089	0.890	0.779-1.018	†
x7	0.153	0.068	2.239	0.025	1.165	1.019-1.331	$*$
x8	0.104	0.068	1.518	0.129	1.109	0.970-1.268	
x9	0.089	0.078	1.136	0.256	1.093	0.937-1.275	
x10	-0.085	0.078	-1.085	0.278	0.919	0.788-1.071	
x11	-0.005	0.068	-0.069	0.945	0.995	0.871-1.137	
x12	0.028	0.068	0.418	0.676	1.029	0.900-1.176	
x13	-0.039	0.068	-0.571	0.568	0.962	0.842-1.099	
x14	-0.014	0.068	-0.199	0.842	0.987	0.863-1.127	
x16	0.048	0.078	0.612	0.540	1.049	0.900-1.224	
x17	-0.012	0.078	-0.152	0.879	0.988	0.848-1.152	



(a) (b)

FIGURE 3. (a) LASSO coefficient path and (b) feature importance

selection and model interpretability, facilitating robust identification of key risk factors for screening purposes. The left plot illustrates the regularization path of LASSO across varying $\log(\lambda)$ values, while the right plot shows the final selected features with non-zero coefficients, indicating their predictive importance for osteoporosis classification.

Feature selection was performed using four complementary approaches: logistic regression analysis, LASSO regularization, random forest importance ranking, and correlation analysis. Logistic regression revealed that x1 was the strongest predictor (Ad-jOR = 29.16, $p < 0.001$), with x7 also demonstrating statistical significance (p = 0.025), and x6 showing borderline significance. LASSO identified x7, x11, x12, x10, and x18 as relevant features, with x11 and x12 contributing the largest coefficients. Random forest analysis further emphasized the dominance of x1, followed by moderate contributions from x9, x4, x14, x13, and x6. Correlation analysis confirmed a strong association between x1 and the outcome ($r \approx 0.69$), while also highlighting potential multicollinearity among x12, x14, x15, x16, and x18. Based on the combined evidence, x1 and x7 were consistently identified as key predictors across

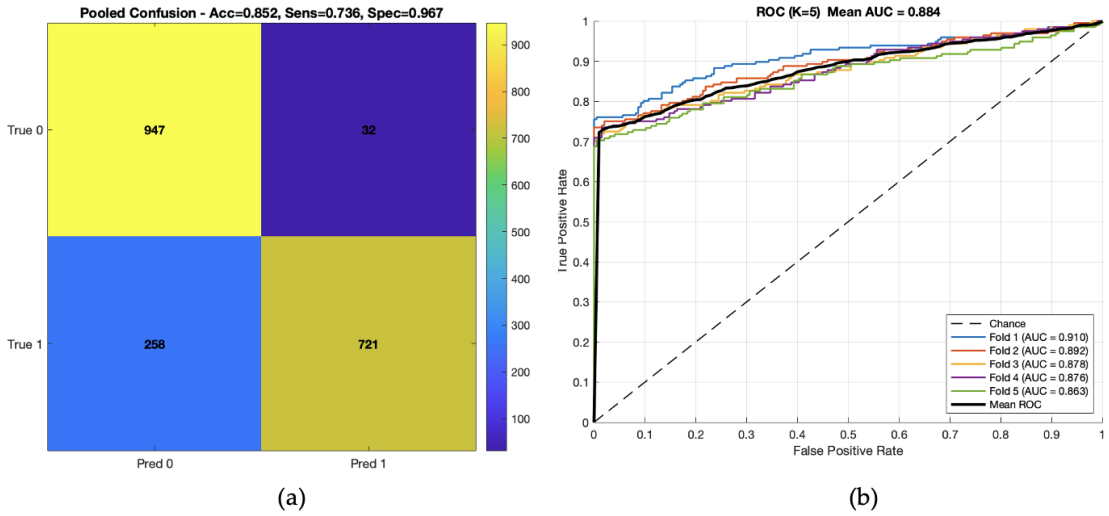


FIGURE 4. Model performance evaluation of random forest. (a) shows the pooled confusion matrix with overall accuracy of 0.852, sensitivity of 0.736, and specificity of 0.967. (b) presents the ROC curve using 5-fold cross-validation, achieving a mean AUC of 0.884

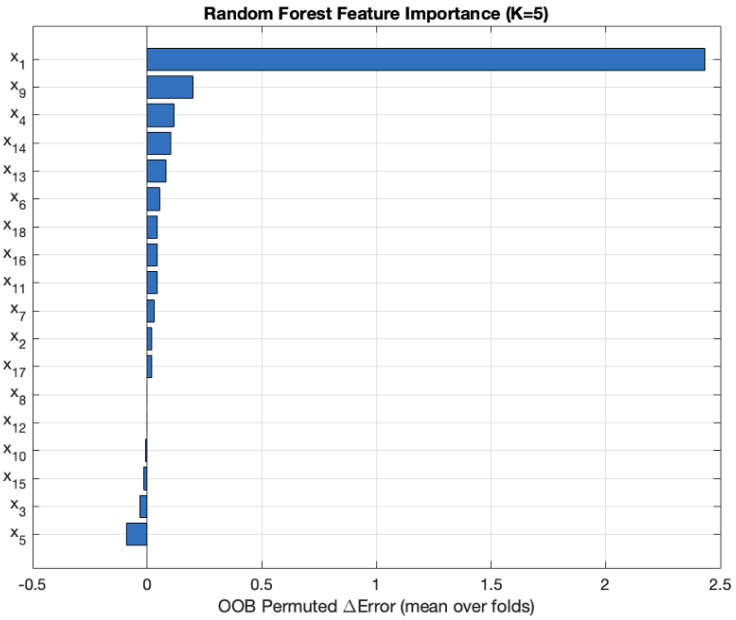


FIGURE 5. Random Forest feature importance (K = 5), showing the mean out-of-bag (OOB) per-muted △Error across folds

multiple methods, while x11, x12, x9, and x6 may provide additional predictive value if managed for multicollinearity. Features with minimal contributions across all methods (e.g., x2, x3, x5, x8, x10, x15, x17) can be excluded to simplify the prediction model without sacrificing performance.

3.3. **Model performance.** Based on the results of feature selection using multiple complementary approaches, including logistic regression, LASSO regularization, Random Forest, and correlation analysis, the predictors were systematically categorized according to their relative importance and stability

TABLE 2. Feature selection summary across four methods. Checkpoints ($\surd$, $\sim$, $\times$) denote strong, partial, or no support

Variable	Logistic Regression	LASSO	Random Forest	Correlation	Overall Decision
x1	$\surd$ (***, AdjOR=29.16)	–	$\surd$ (highest)	$\surd$ (r=0.69)	$\surd$ Keep (key predictor)
x7	$\surd$ (* p=0.025)	$\surd$	$\sim$ (low)	–	$\surd$ Keep
x6	$\sim$ († borderline)	–	$\sim$ (moderate)	–	$\sim$ Possible
x11	$\times$ (NS)	$\surd$	$\sim$ (low)	–	$\sim$ Possible
x12	$\times$ (NS)	$\surd$	$\sim$ (low-mod)	$\sim$ (multicollinear)	$\sim$ Possible
x9	$\times$	–	$\sim$ (moderate)	–	$\sim$ Possible
x4,x4,x13	$\times$	–	$\sim$ (moderate)	$\sim$ (multicollinear)	$\sim$ Optinal
x10,x18	$\times$	$\surd$ (small)	$\times$/ $\sim$	$\sim$ (multicollinear)	$\sim$ Optinal
x2, x3, x5, x8, x15, x16, x17	$\times$	$\times$	$\times$	$\times$	$\times$ Exclude

TABLE 3. Comparative performance of Fine Tree, Optimizable SVM, and Efficient Logistic Regression models across different feature cases. Metrics include training time, accuracy, sensitivity, specificity, F1-score, and AUC

Model	Feature	Training Time (sec)	Accuracy	Sensitivity	Specificity	F1-Score	AUC
Fine Tree	Case 1	9.08	85.6	84.5	86.8	85.5	0.0926
	Case 2	1.78	87.7	83.6	91.8	87.2	0.9151
	Case 3	115.39	89.0	83.3	94.7	88.3	0.9152
	Case 4	277.73	91.4	82.8	100	90.6	0.9208
Optimizable SVM	Case 1	529.45	85.3	73.1	97.4	83.3	0.8874
	Case 2	294.04	86.9	84.3	89.5	86.6	0.9030
	Case 3	213.05	90.9	82.4	99.3	90.0	0.9136
	Case 4	508.03	91.4	82.8	100	90.6	0.9085
Efficient Logistic Regression	Case 1	17.28	82.3	77.7	86.8	81.5	0.9013
	Case 2	2.37	82.2	78.0	86.4	81.4	0.9023
	Case 3	181.92	82.9	78.2	87.6	82.1	0.9026
	Case 4	469.11	82.7	78.4	87.0	82.0	0.9036
Neural Network	Case 1	313.91	85.0	72.7	97.3	83.0	0.8860
	Case 2	601.96	85.6	74.1	97.1	83.9	0.8947
	Case 3	848.87	86.6	73.6	99.7	84.5	0.9046
	Case 4	708.03	88.4	77.9	98.8	87.0	0.9158
Optimizable Naïve Bayes	Case 1	134.67	86.0	74.0	98.0	84.1	0.9163
	Case 2	162.66	86.3	75.3	97.2	84.6	0.9179
	Case 3	848.87	86.6	73.6	99.7	84.6	0.9046
	Case 4	487.43	86.0	75.5	96.5	84.4	0.9194

across methods. Variables consistently identified as strong and clinically relevant predictors (e.g., Age [x1] and Medications [x7]) were retained, whereas others with weaker or inconsistent contributions were considered optional or excluded to reduce model complexity and multicollinearity. Accordingly, four feature subsets were constructed to evaluate the trade-off between predictive performance and model parsimony:

Case1: x1-x18

Case2: x1, x4, x6, x7, x9, x10, x11, x12, x13, x14, x18

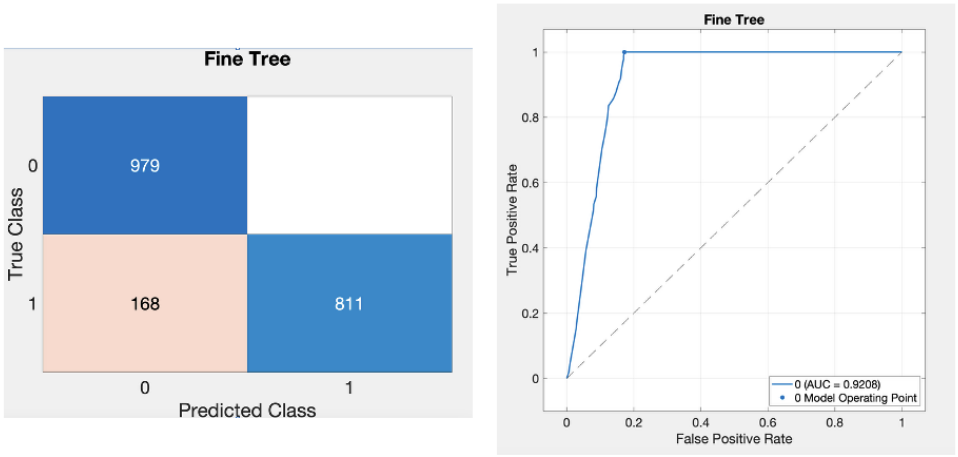
Case3: x1, x6, x7, x9, x11, x12

Case4: x1, x7

From the results in Table 3, the comparative evaluation of each model in terms of both performance and computational cost can be summarized. The trade-off between accuracy, sensitivity, specificity,

TABLE 4. Comparative summary of classification models in terms of performance and computational cost. Optimizable SVM yielded the highest predictive performance, Fine Tree offered a balanced trade-off, while Efficient Logistic Regression minimized training time at the expense of accuracy

Model	Best Case	Accuracy(%)	Sensitivity(%)	Specificity(%)	F1-Score(%)	AUC	Training Time (sec)	Trade-off Summary
Fine Tree	Case 4	91.4	82.8	100	90.6	0.9208	277.73	Good balance between accuracy and specificity with moderate training time.
Optimizable SVM	Case3/4	90.9–91.4	82.4–82.8	99.3–100	90.0–90.6	0.9136–0.9085	213–508	Highest predictive performance, but requires high computational cost.
Efficient Logistic Regression	Case 1	82.3	77.7	86.8	81.5	0.9013	2.37–17.28	Fastest training time, but with lower accuracy and sensitivity.
Neural Network	Case 4	88.4	77.9	98.8	87.0	0.9158	708.03	Strong performance and high AUC, but very high computational cost.
Optimizable Naïve Bayes	Case 4	86.0	75.5	96.5	84.4	0.9194	487.43	High specificity and competitive AUC, but relatively low sensitivity.



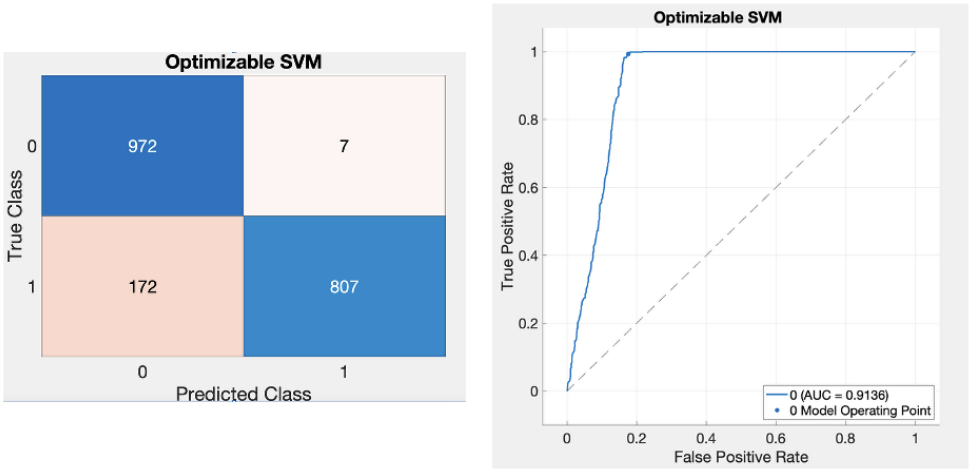
(a) (b)

FIGURE 6. Confusion matrix (a) and receiver operating characteristic (ROC) curve (b) of the Fine Tree model under Case 4 feature selection (x1: Age, x7: Medications). The confusion matrix demonstrates the distribution of correctly and incorrectly classified cases, while the ROC curve shows the model's discriminative ability, achieving an area under the curve (AUC) of 0.9208, indicating excellent predictive performance

and training time highlights the relative strengths and weaknesses of the models, which are further consolidated in the subsequent summary table.

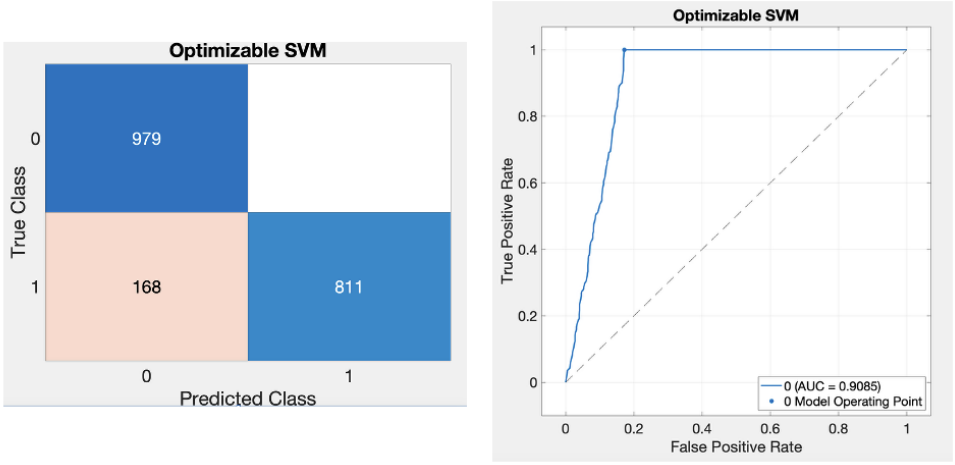
From Table 4 and Figs 6-8, the comparative analysis across multiple machine learning models demonstrates distinct trade-offs between predictive performance and computational cost. Fine Tree and Optimizable SVM consistently achieved the highest accuracy and specificity, particularly in Case 3 and Case 4, with Area Under the Curve (AUC) values exceeding 0.91. These models, however, required substantially longer training times, which may limit their scalability in resource-constrained settings.

Efficient Logistic Regression, while showing relatively lower accuracy and sensitivity, offered the most favorable balance between speed and interpretability, making it suitable for rapid deployment scenarios. Neural Network and Optimizable Naïve Bayes models provided competitive performance,



(a) (b)

FIGURE 7. Performance evaluation of the Optimizable Support Vector Machine (SVM) model (Case 3, feature set = x1, x6, x7, x9, x11, x12). (a) Confusion matrix showing true positives (807), true negatives (972), false positives (7), and false negatives (172). (b) Receiver operating characteristic (ROC) curve with an AUC of 0.9136, indicating strong discrim-inative ability for osteoporosis prediction



(a) (b)

FIGURE 8. Confusion matrix (a) and ROC curve (b) of the Optimizable SVM model using Case 4 (features: age and medications). The model achieved an accuracy of 91.4%, sensitivity of 82.8%, specificity of 100%, F1-score of 90.6%, and an AUC of 0.9085, highlighting strong predictive ability with only two key predictors

with AUC values comparable to Fine Tree and SVM, though they exhibited moderate to high computational demands.

Taken together, the findings indicate that model selection should be guided by application context:

(i) Performance-oriented applications (e.g., high-stakes clinical decision support) would benefit from Fine Tree or Optimizable SVM.
(ii) Cost-sensitive or real-time applications (e.g., large-scale screening or embedded systems) may favor Efficient Logistic Regression.

These trade-offs are systematically summarized in the subsequent checkpoint table, providing a practical guideline for selecting the most appropriate model based on both performance and cost considerations.

4. Discussion

The findings of this study highlight the clinical relevance of predictive modeling for osteoporosis by comparing different machine learning approaches under varying feature selection strategies (Case1–4). Importantly, the progressive reduction of input variables—from the full set of 18 predictors in Case1 to only two highly informative variables (Age [x1] and Medications [x7]) in Case4—provides key insights into the balance between model complexity, interpretability, and predictive accuracy.

From a clinical standpoint, Case1 represents a comprehensive risk profile that in-tegrates demographic, lifestyle, and medical condition variables. While this approach ensures maximal information capture, the inclusion of numerous predictors increases the risk of multicollinearity and complicates practical implementation in real-world healthcare settings. In contrast, Case2 and Case3 streamline the model by focusing on a reduced set of variables (e.g., Age, FamilyHistory, Smoking, Medications, and Prior Fractures), which remain clinically interpretable and more feasible to collect in primary care or screening environments.

Case4 demonstrates that a minimal feature set—Age and Medications—can still achieve competitive performance metrics across models such as Fine Tree and Optimi-zable SVM. This result is particularly valuable for developing rapid, accessible screen-ing tools for older adults, especially in primary care or community health contexts where data availability and resources may be limited. The prioritization of Age as the strongest predictor aligns with established literature emphasizing its role in bone den-sity decline, while the inclusion of Medications reflects the importance of pharmacolog-ical management in lifestyle medicine. These insights directly support novel applica-tions of lifestyle medicine in adults and older adults by enabling targeted interventions that combine early risk identification with preventive and therapeutic strategies.

Regarding model selection, Fine Tree and Optimizable SVM offered the best balance between predictive accuracy and clinical interpretability, although the latter required substantially higher computational resources. Efficient Logistic Regression provided the fastest training times, but at the expense of reduced sensitivity, which is a critical limitation for identifying high-risk patients. Neural Networks and Naïve Bayes achieved strong AUC and specificity values but required longer training times, making them less suitable for resource-constrained environments.

Clinically, these results suggest that predictive models using reduced feature sets can still maintain strong performance, thereby supporting their translation into screen-ing tools or decision-support systems for older adults. The choice of model should be guided by the specific clinical context: in settings prioritizing sensitivity (e.g., early detection in high-risk populations), models trained under Case1 or Case2 may be more appropriate. Conversely, in environments requiring rapid deployment with limited variables, Case4 demonstrates the potential for streamlined models to retain clinical utility.

Overall, this study underscores the importance of aligning predictive modeling strategies with clinical feasibility and the principles of lifestyle medicine. Reducing the number of input variables without compromising predictive accuracy enhances the po-tential for integration into routine clinical workflows, supporting early intervention, preventive care, and personalized risk management in osteoporosis among adults and older adults.

5. Conclusion

This study demonstrated the potential of machine learning–based predictive models in assessing osteoporosis risk, highlighting the importance of both model selection and feature reduction strategies. By systematically comparing five supervised classifiers—Fine Tree, Optimizable SVM, Efficient Logistic Regression, Neural Network, and Optimizable Naïve Bayes—across four feature selection cases, we showed that reducing the number of input variables did not substantially compromise predictive performance. In particular, the combination of age and medication use (Case4) yielded competitive accuracy and specificity, underscoring the feasibility of developing parsimonious yet clinically effective models. From a clinical perspective, the findings suggest that predictive models can be adapted to different healthcare settings depending on resource availability and clinical priorities. Models with a larger feature set (Case1, Case2) may provide higher sensitivity and comprehensive risk profiling, which is valuable for early detection and high-risk patient identification. Conversely, models with reduced features (Case3, Case4) offer practical advantages in terms of data collection efficiency and re-al-world applicability, particularly in primary care or community-based screening pro-grams where time and resources are limited.

Overall, this work emphasizes the utility of integrating machine learning into osteoporosis risk assessment and highlights the balance between predictive accuracy, interpretability, and computational cost. These insights provide a foundation for future clinical decision-support tools aimed at improving early diagnosis, individualized risk management, and ultimately reducing the burden of osteoporosis-related fractures.

Author Contributions

Writing–Original Draft, W. W.; Methodology, W. C.; Writing–Review and Editing, K. N., P. U. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

This study was conducted in accordance with the Dec-laration of Helsinki, the Belmont report, CIOMS Guideline international conference on Harmo-nization in Good Cinical Pratice or ICH-GCP and with approval from the Ethics Committee and Institutional Review Board of Phayao Hospital (Institutional Review Board (IRB) approval, IRB Number: COA No.1.1/053/68).

Data Availability Statement

The dataset utilized in this study, titled Lifestyle factors influenc-ing osteoporosis, is publicly accessible and can be retrieved from Kaggle (https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis).

Statements and Declarations

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank University of Phayao and Thailand Science Research and Innovation Fund.

References

[1] NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. Osteoporosis prevention, diagnosis, and therapy. *Journal of the American Medical Association*, 285:785–795, 2001.

[2] S. L. Bonnick and L. Shulman. Monitoring osteoporosis therapy: bone mineral density, bone turnover markers, or both. *American Journal of Medicine*, 119:S25–S31, 2006.

[3] O. Johnell and J. Kanis. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporosis International*, 17:1726–1733, 2006.

[4] T. Sözen, L. Özışık, and N. Ç. Başaran. An overview and management of osteoporosis. *European Journal of Rheumatology*, 4:46–52, 2016.

[5] E. Hernlund, A. Svedbom, M. Ivergård, J. Compston, C. Cooper, J. Stenmark, E. V. McCloskey, and J. A. Kanis. Osteoporosis in the European Union: medical management, epidemiology and economic burden: a report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA). *Archives of Osteoporosis*, 8:Article ID 136, 2013.

[6] K. Limlomwongse. Prevalence of osteopenia and osteoporosis in Thai women. *Menopause*, 8:65–69, 2001.

[7] C. J. Crandall. Risk assessment tools for osteoporosis screening in postmenopausal women: a systematic review. *Current Osteoporosis Reports*, 13:287–301, 2015.

[8] J. A. Kanis, O. Johnell, A. Odén, H. Johansson, and E. V. McCloskey. FRAX™ and the assessment of fracture probability in men and women from the UK. *Osteoporosis International*, 19:385–397, 2008.

[9] F. Cosman, S. J. de Beur, M. S. LeBoff, E. M. Lewiecki, B. Tanner, S. Randall, and R. Lindsay. Clinician's guide to prevention and treatment of osteoporosis. *Osteoporosis International*, 25:2359–2381, 2014.

[10] X. Jiang, M. Gruner, F. Trémollieres, W. Pluskiewicz, E. Sornay-Rendu, P. Adamczyk, and P. F. Schnatz. Diagnostic accuracy of FRAX in predicting the 10-year risk of osteoporotic fractures using the USA treatment thresholds: a systematic review and meta-analysis. *Bone*, 99:20–25, 2017.

[11] J. A. Kanis, N. Burlet, C. Cooper, P. D. Delmas, J.-Y. Reginster, F. Borgstrom, and R Rizzoli. European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporosis International*, 19:399–428, 2008.

[12] W. D. Leslie, S. Morin, L. M. Lix, H. Johansson, A. Odén, E. McCloskey, and Manitoba Bone Density Program. Fracture risk assessment without bone density measurement in routine clinical practice. *Osteoporosis International*, 23:75–85, 2012.

[13] S. H. Ralston and A. G. Uitterlinden. Genetics of osteoporosis. *Endocrine Reviews*, 31:629–662, 2010.

[14] C. M. Weaver, D. D. Alexander, C. J. Boushey, B. Dawson-Hughes, J. M. Lappe, M. S. LeBoff, S. Liu, A. C. Looker, T. C. Wallace, and D. D.Wang. Calcium plus vitamin D supplementation and risk of fractures: an updated meta-analysis from the National Osteoporosis Foundation. *Osteoporosis International*, 27:367–376, 2016.

[15] B. M. Tang, G. D. Eslick, C. Nowson, C. Smith, and A. Bensoussan. Use of calcium or calcium in combination with vitamin D supplementation to prevent fractures and bone loss in people aged 50 years and older: a meta-analysis. *Lancet*, 370:657–666, 2007.

[16] D. Stacey, F. Légaré, K. Lewis, M. J. Barry, C. L. Bennett, K. B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, R. Thomson, and L. Trevena. Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews*, 4:Article ID CD001431, 2017.

[17] E. S. Siris, R. Adler, J. Bilezikian, M. Bolognese, B. Dawson-Hughes, M. J. Favus, S. T. Harris, S. M. J. de Beur, S. Khosla, N. E. Lane, R. Lindsay, A. D. Nana, E. S. Orwoll, K. Saag, S. Silverman, and N. B.Watts. The clinical diagnosis of osteoporosis: a position statement from the National Bone Health Alliance Working Group. *Osteoporosis International*, 25:1439–1443, 2014.

[18] L. Shepstone, E. Lenaghan, C. Cooper, S. Clarke, R. Fong-Soe-Khioe, R. Fordham, N. Gittoes, I. Harvey, N. Harvey, A. Heawood, R. Holland, A. Howe, J. Kanis, T. Marshall, T. O'Neill, T. Peters, N. Redmond, D. Torgerson, D. Turner, E. McCloskey, and SCOOP Study Team. Screening in the community to reduce fractures in older women (SCOOP): a randomised controlled trial. *Lancet*, 391:741–747, 2018.

[19] L. Shepstone, R. Fordham, E. Lenaghan, I. Harvey, C. Cooper, N. Gittoes, T. J. Peters, T. O'Neill, D. Torgerson, R. Holland, A. Howe, T. Marshall, J. A. Kanis, and E. McCloskey. A pragmatic randomised controlled trial of the effectiveness and cost-effectiveness of screening older women for the prevention of fractures: rationale, design and methods for the SCOOP study. *Osteoporosis International*, 23:2507–2515, 2012.

[20] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2:Article ID e000101, 2017.

[21] G. Bonaccorsi, M. Giganti, M. Nitsenko, G. Pagliarini, G. Piva, and G. Sciavicco. Predicting treatment recommendations in postmenopausal osteoporosis. *Journal of Biomedical Informatics*, 118:Article ID 103780, 2021.

[22] A. Kulkarni. Lifestyle factors influencing osteoporosis [Dataset]. *Kaggle*. Available online: https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis (accessed on 20 August 2025).

[23] J. T. Hancock and T. M. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Dat*, 7:Article ID 28, 2020.

[24] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carr´e, J. R. G. Marqu´ez, B. Gruber, B. Lafourcade, P. J. Leit˜ao, T. M¨unkem¨uller, C. Mcclean, P. E. Osborne, B. Reineking, B. Schr¨oder, A. K. Skidmore, D. Zurell, and S. Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36:27–46, 2013.

[25] M. Kuhn and K. Johnson. Applied Predictive Modeling, volume 26. Springer, New York, USA, 2013.

[26] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. Applied Logistic Regression. John Wiley & Sons, Hoboken, New Jersey, USA, 2013.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288, 1996.

[28] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[29] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, USA, 2017.

[30] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[31] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[32] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[33] Z. N. A. M. S. Alden and O. Ata. A comprehensive analysis and performance evaluation for osteoporosis prediction models. *PeerJ Computer Science*, 10:Article ID e2338, 2024.

[34] G. Naidu, T. Zuva, and E. M. Sibanda. A review of evaluation metrics in machine learning algorithms. In R. Silhavy, P. Silhavy, editors, Proceedings of the Computer Science On-line Conference (CSOC), pages 15–25, Cham, Switzerland, 25–28 April 2023, Springer, Cham, Switzerland.