ISSN 2723-6471

Osteoporosis Detection Using a Combination of Recursive Feature Elimination and Naive Bayes Classifier with Rule-Based Chatbot Testing

Enny Itje Sela[1,*], Rianto[2], Afwan Anggara[3], Wahyu Sri Utami[4]

[1]*Information Technology Study Program, University Technology of Yogyakarta*

[2,3,4]*Science and Technology Department, University Technology of Yogyakarta*

(Received: August 21, 2024; Revised: September 14, 2024; Accepted: September 21, 2024; Available online: October 20, 2024)

Abstract

Osteoporosis is a condition characterized by reduced bone mass and density, increasing the risk of fractures. Early detection relies on patient awareness and proactive health management. Despite advances in technology, patient independence and awareness remain critical for early diagnosis. A rule-based chatbot tool can assist by helping patients screen their bone health. The chatbot provides automated recommendations, offering an alternative to traditional hospital visits. This study presents a rule-based chatbot designed to detect osteoporosis, using Recursive Feature Elimination (RFE) combined with the Naïve Bayes Classifier (NBC). Machine learning is integrated to enhance the chatbot's ability to identify early signs of osteoporosis. The model's performance is compared to other feature selection methods, such as Principal Component Analysis (PCA), and machine learning algorithms like Deep Learning, Support Vector Machine (SVM), and Logistic Regression. The dataset used includes public data sets for training and validation, as well as data from the Yogyakarta Health Office for predictions. Research phases include normalization, data encoding, feature selection, training, validation, and prediction. The chatbot implements text preprocessing techniques, such as tokenization, stop word removal, and feature extraction, alongside normalization and encoding of numeric data. The prediction stage determines if the patient has a positive or negative osteoporosis status. Validation results show the RFE-NBC model is particularly effective for osteoporosis detection, offering a balanced performance in identifying both positive and negative cases. Additionally, this model served as the foundation for creating a rule-based chatbot designed to detect osteoporosis. Based on the set of testing metrics using chatbot, the model demonstrates strong overall performance, with a good balance between identifying positive and negative instances.

Keywords: Osteoporosis, Chatbot, Naïve Bayes, Recursive Feature Elimination, Machine-Learning

1. Introduction

Osteoporosis is a bone health condition marked by decreased bone mass and density, which increases the risk of fractures [1], [2]. Often asymptomatic in its early stages, early diagnosis is essential to prevent further complications [3], [4], [5]. The success of early osteoporosis detection heavily relies on patient awareness and active involvement in maintaining bone health. By emphasizing patient empowerment, individuals can become more engaged in managing their health [6]. Despite technological advancements in osteoporosis detection, enhancing patient independence and awareness remains crucial. Strengthening these aspects can improve understanding of bone health and promote early detection of osteoporosis [7], [8].

One solution to address this issue is the development of a chatbot tool to assist patients in screening their bone health. The use of chatbots in health services represents an innovative advancement in information technology. Chatbots can effectively enhance health awareness, provide information, and even facilitate early disease detection [9]. These artificial intelligence programs simulate human conversations through text or voice messages on messaging platforms, websites, or mobile applications [10]. The use of chatbots in healthcare offers several distinct advantages over traditional diagnostic tools. Chatbots provide natural language interactions, allowing patients to describe their symptoms more easily without needing medical jargon. They are available 24/7, offering instant support outside of standard medical hours, unlike traditional tools that require professional operation and scheduled appointments.

***Corresponding author: Enny Itje Sela (ennysela@uty.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i4.409

Additionally, chatbots are scalable, handling multiple users simultaneously, while traditional diagnostics are resource intensive. They serve as decision-support tools, offering advice based on patient input, though not replacing doctors. Moreover, chatbots can track patient data over time for personalized care, whereas traditional tools are typically used for one-time assessments. Rule-based chatbots, in particular, are designed with specific conversation flows based on predetermined rules [11]. They respond to queries by matching input with predefined keywords [12].

Many studies have successfully applied machine learning techniques to detect osteoporosis, utilizing various data types such as dental images [1], [4], [13], hip images [14], [15], spine images [16], [17], and anthropometric features [18]. However, these studies did not incorporate chatbots in their testing or prediction processes. This study addresses this gap by developing a rule-based chatbot for osteoporosis detection using several machine learning techniques, namely Deep Learning (DL), Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), and Logistic Regression (LR), combined with Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA). The Naïve Bayes Classifier performs well even with limited training data due to its assumption of feature independence, which simplifies the model [19]. Deep Learning is well-suited for tasks involving large and complex datasets [20]. SVM perform effectively with smaller datasets and scenarios where there are significant margins between classes [21], Logistic Regression is best for straightforward classification tasks and when model interpretability is crucial [22].

Feature selection is crucial before training with machine learning. It helps identify dominant features, enhances model performance, reduces overfitting, speeds up training, and aids in model interpretation. The feature selection method used in this study is RFE. RFE works by iteratively removing the least important features according to criteria set by the model, thereby focusing on the most relevant features and significantly improving model predictions [23], [24]. PCA is used for feature selection because it helps reduce dimensionality while retaining as much variance (information) as possible. Unlike methods like RFE, PCA does not necessarily select the original features but rather creates new combinations of them. Several studies on osteoporosis detection have successfully used PCA for feature selection [25], [26].

This study proposed a model for developing a rule-based chatbot for the early detection of osteoporosis, employing RFE for feature selection and the Naïve Bayes Classifier for the learning process compared to other feature selection and machine learning algorithms. Using a rule-based chatbot enables patients to easily access information, conduct self-examinations, and understand osteoporosis risk factors. A rule-based chatbot is well-suited for osteoporosis detection because it operates on predefined rules and logic, making it reliable for handling specific symptoms and risk factors associated with the disease.

2. Literature Review

Table 1 highlights several relevant studies related to this research, which have utilized machine learning techniques for osteoporosis detection, achieving validation accuracy rates exceeding 80%. However, none of these studies incorporated chatbots for prediction. Study [5] explored the use of deep learning to diagnose osteoporosis from hip radiographs and evaluated whether incorporating clinical data improved diagnostic performance compared to using image data alone. Osteoporosis was assessed using five convolutional neural network (CNN) models, with EfficientNet-b3 achieving the highest accuracy. Study [15] aimed to predict osteoporosis using simple hip radiography with a deep learning algorithm. A deep neural network (DNN) model was developed based on VGG16, enhanced with a nonlocal neural network. The final DNN model achieved an overall accuracy of 81.2%. Mookiah et al. [27] differentiated between healthy individuals and those with osteoporotic fractures by using texture features extracted from CT images, achieving a classification accuracy of 83%. This study demonstrated the feasibility of opportunistic osteoporosis screening through CT image texture analysis. Study [28] examined the effectiveness of various machine learning (ML) techniques in classifying postmenopausal Thai women with osteoporosis. The study compared pre-processed and original data to assess the performance of different ML methods. The results indicated that different ML algorithms, when combined with pre-processing techniques, produced varied outcomes. The Wrapper Bayesian Network method, applied to the Neural Network model, achieved the highest accuracy of 83.8%. Study [29] designed multiple heterogeneous machine learning frameworks to predict the risk of osteoporosis. An open-source dataset of 1,493 patients, containing bone density, blood, and physical test data, was utilized. The best-performing pipeline used a Forward Feature Selection algorithm followed by a custom multi-level ensemble learning-based stack, achieving an

ISSN 2723-6471

accuracy of 89%. A layer of explainable artificial intelligence (XAI) and feature importance provided interpretability and insight into the classifier's predictions. Kwon et al. [30] developed an ensemble machine learning model to screen for osteoporosis among postmenopausal Korean women. Data from 1,431 patients were used, with 20 features extracted through feature importance and RFE. Three tree-based models—Random Forest (RF), AdaBoost, and Gradient Boosting—were trained, with AdaBoost achieving an accuracy of 82.9%.

On the other hand, several studies successfully implemented self-learning chatbots for diseases other than osteoporosis. Deshpande et al. [31] designed a self-harm classifier that uses a user's responses to a chatbot to predict whether the response indicates intent for self-harm. A sentiment analysis classifier was trained using Twitter data, and the results were combined with another model to enhance performance. The best results were achieved with an LSTM-RNN classifier using BERT encoding, reaching an accuracy of 92.13%. Study [32] used a COVID-19 information dataset to evaluate the proposed methodology. The pandemic was accompanied by an "infodemic" of fake news, and the study aimed to measure accuracy, effectiveness, efficiency, and satisfaction. The Naive Bayes model achieved the highest accuracy at 88.12%. Gao et al. [33] demonstrated the potential of various readability metrics as features to predict the popularity of chatbots. Their study revealed that highly popular and unpopular chatbots have significant differences in readability scores, suggesting that readability metrics can be a valuable indicator of user interest in chatbot adoption.

Chakraborty et al. [34] proposed a medical chatbot that handles human interaction and predictive tasks using a MLP. This current study introduces a model for osteoporosis detection utilizing a rule-based chatbot, with its performance powered by RFE and the NBC.

Table 1. Literature Review

Researchers	Domain	Chatbot	Method of detection	Accuracy
Yamamoto, et al [5]	Osteoporosis classification from hip image	No	DL, EfficientNetb3	86,73%
Mookiah, et al [27]	CT scan image	No	GLCM, SVM	83%
Thawnashom, et.al [28]	Classifying postmenopausal osteoporosis Thai patients	No	NN, Bayesian Network	83.8%
Jang, et.al [15]	Prediction osteoporosis from hip radiography	No	DL VGG 15	81.2%
Khanna et.al [29]	Osteoporosis risk prediction	No	Forward feature selection and XAI	89%
Kwon et al. [30]	Screening OP among postmenopausal Korean women	No	RFE, AdaBost	82.90%
Deshpande et al. [31]	Self-Harm Detection for Mental Health Chatbots	Yes	LSTM, RNN, classifier, BERT encoding	92.13%
Ghaleb et al. [32]	Development and evaluation of a microservice-based virtual assistant for chronic patients' support	Yes	NLP	88.12%
Gao et al. [33]	Computational approach to extracting features and training models that make a priori prediction about chatbots' popularity	Yes	NLP	77.36%
Chakraborty etc. [34]	An AI-Based Medical Chatbot Model for Infectious Disease Prediction	Yes	MLP	94.32%

3. Methodology

This research is conducted in a structured manner according to the research steps (figure 1). The first stage is collecting relevant data for osteoporosis detection. This study utilizes two primary data sources: a Kaggle dataset, which consists of 16 features and includes 1,958 records, with the features detailed in table 2, and survey data collected by researchers in Yogyakarta, Indonesia, involving 43 participants. This survey was conducted between January 2023 and August 2024, with permission obtained from the local Posyandu (integrated health post) in Yogyakarta. The subjects of the survey were women aged 30 to 50 years. The survey collected information on 17 features relevant to osteoporosis, including age, gender, hormonal changes, family history, race/ethnicity, weight, height, calcium intake, vitamin d

 ISSN 2723-6471

intake, physical activity, smoking, alcohol consumption, medical conditions, medications, prior fractures, and osteoporosis status. Additionally, the researchers utilized ultrasonography as part of the assessment process to gather more detailed information on bone health. Out of the 43 data points collected, 15 belong to the osteoporosis-positive class, while the remaining are categorized as osteoporosis-negative.

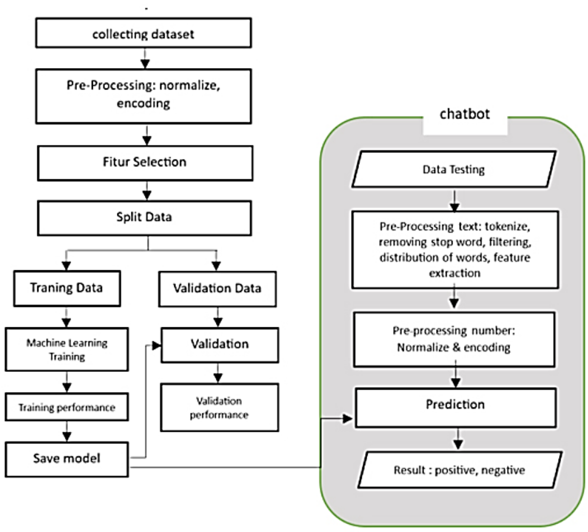

Figure 1. The steps of research

The collected data requires further processing to make it suitable for the machine learning model. This process, known as data normalization, involves adjusting the feature values to a consistent scale. This ensures that the model can learn more effectively, as it is not biased towards features with larger scales. The method used for this is the standard scaler, a mean-based scaling method. It adjusts the mean to 0, but it’s important to note that the StandardScaler is sensitive to outliers, as they can significantly affect the mean.

Table 2. Features of dataset

Features	Indicates
Id	Unique identifier
Age	The age of the individual in years
Gender	The gender of the individual: "Male" or "Female"
Hormonal Changes	indicates whether the individual has undergone hormonal changes, particularly related to menopause: "Postmenopausal" for females or "Normal" for individuals who haven't experienced significant hormonal changes.
Family History	Indicates whether there is a family history of osteoporosis or fractures: "Yes" or "No".
Race/Ethnicity	The race or ethnicity of the individual: "Caucasian", "African American", "Asian", etc.
BMI	The body mass index status of the individual: "Normal" or "Underweight" or “overweight”
Calcium Intake	The level of calcium intake in the individual's diet: "Low" or "Adequate"
Vitamin D Intake	The level of vitamin D intake in the individual's diet: "Insufficient" or "Sufficient".
Physical Activity	Indicates the level of physical activity of the individual: "Sedentary" for low activity levels or "Active" for regular exercise
Smoking	Indicates whether the individual is a smoker: "Yes" or "No".
Alcohol Consumption	Indicates the level of alcohol consumption by the individual: "None" for non-drinkers or "Moderate" for moderate drinkers
Medical Conditions	Any existing medical conditions that the individual may have: "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: "Yes" or "No".
Osteoporosis	The target variable indicates the presence or absence of osteoporosis. This is the variable that we want to predict using machine learning algorithms. It can be "1" for presence or "0" for absence of osteoporosis.

The standard scaler formula is presented in (1).

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}} \quad (1)$$

In addition to normalization, an encoding process, changing categorical data into numeric form, is also carried out. The encoding method used is the one-hot encoding technique. In one-hot encoding, each unique category is represented by a binary vector whose length equals the number of categories. Each element in the vector has a value of 0, except for one element which has a value of 1, which indicates the presence of the category. This technique eliminates the false assumption that there is an ordinal relationship among the categories. The feature selection stage involves selecting the most relevant and significant features for osteoporosis detection. Irrelevant features can be ignored to improve model performance. One popular feature selection technique is RFE [24].

The algorithm for RFE [24] begins by assuming there are n features in the dataset, and m is the desired number of features to select. The first step is to train a random forest machine learning model using all the features. Feature importance or coefficients are then obtained based on the change in Mean Squared Error (MSE) when a specific feature is removed. This change is represented by the formula (2)

$$\Delta MSE(f_i) = MSE_{with\ f_i} - MSE_{withoutf_i} \quad (2)$$

Next, the features are ranked by sorting them according to the absolute values of their importance based on the model coefficients. The vector of feature importance is denoted as w=[w1,w2,...,wn] The least important feature is then eliminated, which means the feature with the smallest absolute value |Ik|, is removed from the dataset. Finally, the process is repeated until only mmm features remain. This approach helps in selecting the most important features for the model.

The data was then split into training and validation sets. In this study, the training-to-validation ratios were set at 80%:20%, 85%:15%, 70%:30%, 75%:25%, 65%:35%, and 60%:40%. The NBC was used to implement both the training and validation processes.

Here are the steps for the NBC algorithm [19] The steps for the NBC algorithm begin by calculating the mean (μik) and variance (σ2ik), for each feature xi in class ck. The mean is computed as (3). The variance for each feature in class ck is computed as variance (4):

$$\mu_{ik} = \frac{1}{N_k} \sum_{j \in c_k} x_{ij} \quad (3)$$

$$\sigma_{ik}^2 = \frac{1}{N_k} \sum_{j \in c_k} (x_{ij} - \mu_{ik})^2 \quad (4)$$

where Nk the number of samples in class ck. Next, the prior probability for each class is calculated by dividing the number of samples in class ck by the total number of samples N (5)

$$P(c_k) = \frac{N_k}{N} \quad (5)$$

Following this, the likelihood for each feature xi in class ck is calculated using the Gaussian (normal) distribution. The likelihood is computed as (6):

$$P(x_{i,}c_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right) \tag{6}$$

Here, μik is the mean and σik2 is the variance of feature xi for class ck. Finally, the probability for each class ck is calculated using Bayes' Theorem (7):

$$P(c_k|x) \propto P(c_k)\prod_{i=1}^{n} P(x_i|c_k) \tag{7}$$

where x = (x1, x2, …, xn) is features vector. This provides the classification probability for class ck based on the features in the dataset (8):

$$\hat{c} = \mathrm{arc\,max}\, P(c_k|x) \tag{8}$$

where $\hat{c}$ is the predicted class, P(ck) is the prior probability, and P (xi | ck) is the likelihood for each feature xi.

Next, the model's performance is evaluated using a confusion matrix. The trained and validated model is saved to a file, serving as the knowledge base for subsequent predictions. The testing phase is conducted through an interactive chatbot, which takes user input, processes the data, and delivers prediction results. During this phase, text input is processed through tokenization, stop word removal, and filtering, while numeric data is normalized and encoded similarly to the initial data preprocessing steps. The trained model then uses the processed data to make predictions, determining whether the user is indicated as positive or negative for osteoporosis.

4. Results and Discussion

4.1. Results

In the preprocessing stage, encoding is applied to 14 features: Gender, Hormonal Changes, Family History, Race/Ethnicity, BMI, Calcium Intake, Vitamin D Intake, Physical Activity, Smoking, Alcohol Consumption, Medical Conditions, Medications, and Prior Fractures. Simultaneously, the Age feature undergoes standard scaler normalization. During feature selection with the RFE algorithm, the 14 features are reduced to 10: Age, Gender, Hormonal Changes, BMI, Calcium Intake, Vitamin D Intake, Physical Activity, Smoking, Medications, and Prior Fractures. The four features eliminated are: Alcohol Consumption, Diet Type, Family History of Osteoporosis, and Menopause Age. Excessive alcohol consumption can impact bone density, its direct contribution may be weaker in comparison to other features like age, hormonal changes, or prior fractures. Alcohol's effects might be more secondary or indirectly captured by other lifestyle factors like BMI or smoking. The overall diet type might be a broad category and less specific than calcium or vitamin D intake, which are directly linked to bone health. Thus, it may not add much additional predictive value once those key nutritional factors are already considered. Though family history can influence the likelihood of osteoporosis, it might have been redundant or highly correlated with other factors like age, hormonal changes, or gender, making it less important in improving the prediction model. Menopause itself is captured under "Hormonal Changes," and menopause age might not significantly improve the model beyond general hormonal changes or other stronger predictors like age and gender.

In essence, these eliminated features likely had less unique predictive power or were redundant compared to the retained features, which more directly affect osteoporosis risk. RFE prioritizes features that contribute the most to improving the model's performance. Data for these 10 features is stored and used as the rule base for training and validating the Naïve Bayes Classifier. Training is conducted with three different data splitting ratios: 90%:10%, 85%:15%, 80%:20%, 75%:25%, 70%:30%, 65%:35%, and 60%:40%. Simulation results show that the best validation accuracy is achieved with a 65%:35% ratio, as illustrated by the confusion matrix in figure 2. This figure indicates that True Positives (TP) = 273, True Negatives (TN) = 317, False Positives (FP) = 20, and False Negatives (FN) = 76. the accuracy, precision, recall, specificity, and F1 Score are 86.01%%, 93.17%, 78.22%, 94.07%, and 85.05%, respectively.

The model is then saved as a `model.pkl` file and utilized for predictions based on the researcher’s survey data.

 ISSN 2723-6471

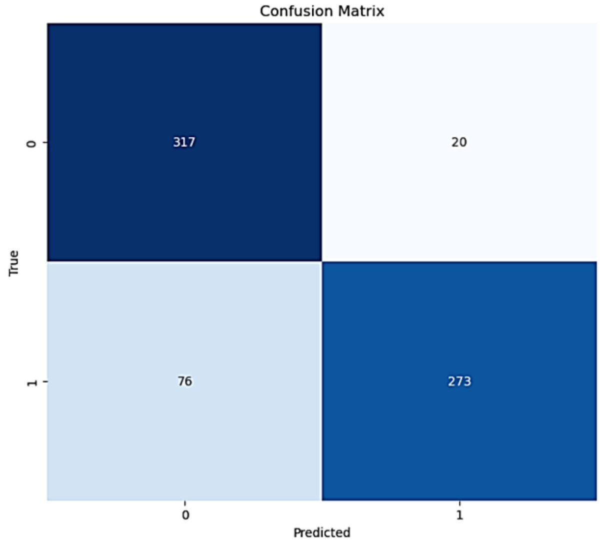

Figure 2. Confusion matrix's validation

Figure 3 illustrates the osteoporosis prediction process for one subject, who received a positive osteoporosis status result. In contrast, figure 4 demonstrates the model's prediction of a negative osteoporosis status based on the input provided by subject 2.

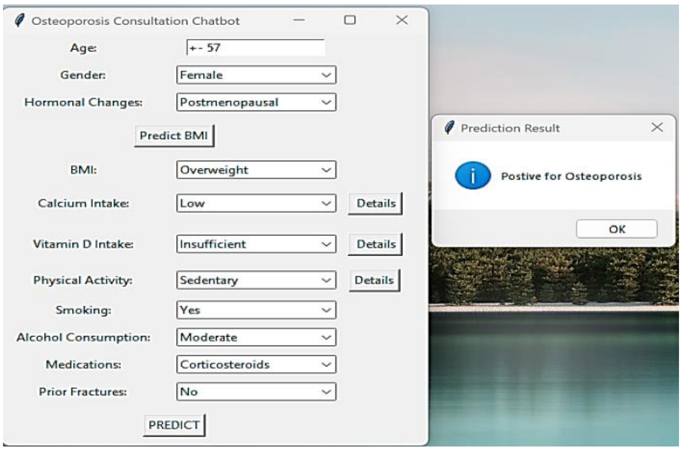

Figure 3. Positive result status of osteoporosis

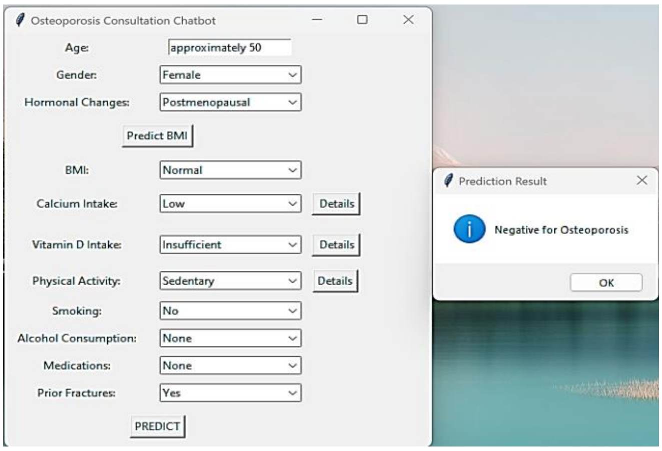

Figure 4. Negative result status of osteoporosis

This model enhances user interaction by accepting flexible and approximate inputs, allowing users to provide estimates rather than requiring exact values. For example, when inputting numeric features such as age, body weight, or height, users can use terms like "about," "around," or "+/-" to indicate approximate figures. This flexibility improves usability, as users may not always know or remember exact details, especially in healthcare contexts where rough estimates (e.g., "around 60 years old" or "about 70 kg") are common. By interpreting these approximations accurately, the model can maintain accuracy in its predictions while offering a more accessible and user-friendly interface. This feature reduces friction in user interactions, helping non-expert users feel more comfortable providing their data, which ultimately improves engagement and the overall accuracy of the system's osteoporosis detection process.

Out of the 43 data tested using the chatbot, the results are as follows: True Positives (TP) = 12, True Negatives (TN) = 25, False Positives (FP) = 3, and False Negatives (FN) = 3. The evaluation metrics obtained are accuracy 86.05%, precision 80.00%, recall 80.00%, specificity 89.29%, and F1 Score 80.00%.

The application also includes a BMI calculator (see figure 5), which can determine BMI categories (Underweight, Normal, Overweight) based on body weight and height. Additionally, the application provides information on foods high in calcium and sufficient in vitamin D, as well as physical activities categorized as active (see figure 6). These features assist users in evaluating their calcium, vitamin D, and physical activity inputs.

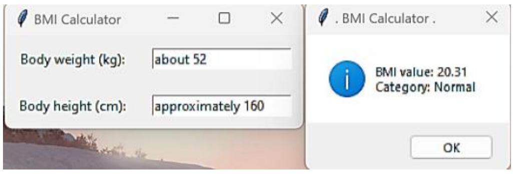

Figure 5. BMI's calculator

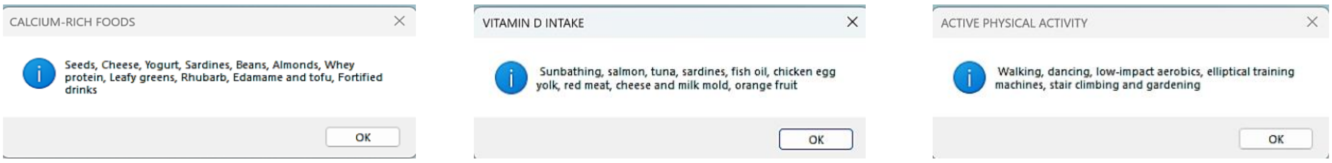

Figure 6. Detail of Calcium-rich food, vitamin D intake and active physical activity

4.2. Discussion

Osteoporosis can lead to bone fractures, commonly affecting the hip, wrist, and spine [16]. While osteoporosis itself is rarely fatal, the complications from fractures, especially hip fractures, can be severe. These complications may include prolonged immobilization, a reduced quality of life, and postoperative issues, which can significantly increase the risk of death, particularly among older adults [23].

This study has developed an osteoporosis prediction model using a combination of RFE for feature selection and the NBC for the learning process. Other selection feature PCA and classification techniques, such as DL, SVM, and LR, were also employed. A comparison of the validation performance of these four techniques is presented in table 3 and figure 7. Based on the experimental results, the validation outcomes using the four machine learning algorithms indicate that the NBC produced the best performance at a 65%:35% ratio. Accuracy (86.01%) indicates that the model is correctly classifying most cases overall. Precision (93.17%) shows the model is highly accurate in predicting positive cases, with few false positives. Recall (78.22%) is somewhat lower, suggesting that while the model predicts positives accurately, it misses a portion of the actual positive cases (false negatives). High Specificity (94.07%) means the model is very effective at identifying true negatives, with few false positives. In a healthcare context like osteoporosis detection, this suggests the model is very good at correctly identifying patients who do not have the condition. This reduces the risk of false positives, meaning fewer patients are incorrectly flagged by having osteoporosis. This is crucial in healthcare because false positives could lead to unnecessary treatment or anxiety for patients. F1 Score (85.05%) balances precision and recall, reflecting strong overall performance but emphasizing that recall could be improved. Additionally, cross-validation has been applied to the dataset model, with k values chosen as factors of the number of dataset records, specifically 11, 22, 89, and 179. The accuracy reached 85.75% at k=179 with a model combining RFE and SVM. However, this accuracy is lower compared to the accuracy achieved using the split ratio method.

NBC can outperform methods due to its simplicity and efficiency [35]. Its assumption of feature independence fits well with osteoporosis risk factors, which may contribute independently to the condition. NBC also performs effectively with smaller datasets, requires less computational power, and handles categorical data, common in medical datasets with ease [36]. Additionally, it offers interpretability, crucial in healthcare and remains robust even when irrelevant features are present [37], making it a practical choice for quick, accurate predictions in osteoporosis detection.

Table 3. Comparison of Validation Results

% Split	**RFE**					**PCA**				
	Accuracy	***Precision***	***Recall***	***Specificity***	***F1 score***	***Accuracy***	***Precision***	***Recall***	***Specificity***	***F1 score***
					DL					
60-40	82.40%	92.62%	70.41%	94.39%	80.00%	83.09%	92.19%	72.30%	93.88%	81.05%
65-35	83.67%	94.60%	71.42%	95.91%	81.39%	84.55%	94.83%	73.63%	95.84%	82.90%
70-30	83.16%	95.14%	69.90%	96.43%	80.59%	82.14%	85.36%	78.88%	85.61%	81.99%
75-25	83.47%	96.06%	69.80%	97.14%	80.85%	80.00%	84.51%	73.47%	86.53%	76.60%

 ISSN 2723-6471

80-20	83.93%	95.24%	71.43%	96.43%	81.63%	80.95%	91.74%	68.03%	93.88%	78.12%
85-15	81.63%	94.28%	67.34%	95.92%	78.57%	80.61%	93.27%	65.99%	95.24%	77.29%
90-10	81.12%	92.96%	67.35%	94.90%	78.11%	75.00%	81.25%	65.65%	84.54%	72.62%
SVM										
60-40	84.57%	95.15%	73.50%	96.09%	82.93%	83.55%	93.40%	72.19%	94.90%	81.44%
65-35	84.55%	94.83%	73.63%	95.84%	82.90%	82.36%	93.02%	69.97%	94.75%	79.87%
70-30	84.52%	94.54%	74.26%	95.44%	83.18%	81.97%	92.34%	69.73%	94.22%	79.46%
75-25	85.51%	100.00%	72.49%	100.00%	84.04%	81.22%	91.80%	68.57%	93.88%	78.50%
80-20	85.20%	100.00%	70.85%	100.00%	82.94%	82.14%	93.75%	68.88%	95.41%	79.41%
85-15	83.33%	100.00%	67.55%	100.00%	80.63%	80.61%	93.27%	65.99%	95.24%	77.29%
90-10	79.59%	1.00%	59.60%	1.00%	74.68%	81.12%	96.92%	64.29%	97.96%	77.30%
NBC										
60-40	85.71%	93.11%	77.75%	94.01%	84.74%	83.42%	91.46%	73.72%	93.11%	81.64%
65-35	86.01%	93.17%	78.22%	94.07%	85.05%	83.09%	92.19%	72.30%	93.88%	81.05%
70-30	85.71%	92.94%	78.22%	93.68%	84.95%	84.21%	90.95%	71.77%	92.86%	80.23%
75-25	85.51%	93.90%	77.52%	94.40%	84.93%	81.63%	90.58%	70.61%	92.65%	79.36%
80-20	84.69%	92.64%	75.88%	93.78%	83.43%	81.63%	89.74%	71.43%	91.84%	79.55%
85-15	82.65%	93.10%	71.52%	94.40%	80.90%	80.95%	91.74%	68.03%	93.88%	78.12%
90-10	78.57%	92.54%	62.63%	94.85%	74.70%	81.12%	92.96%	67.35%	94.90%	78.11%
LR										
60-40	82.14%	84.39%	79.75%	84.64%	82.01%	81.38%	85.76%	75.26%	87.50%	80.16%
65-35	81.78%	83.94%	79.36%	84.27%	81.59%	80.90%	85.57%	74.34%	87.46%	79.56%
70-30	82.14%	85.36%	78.88%	85.61%	81.99%	80.44%	85.10%	73.81%	87.07%	79.05%
75-25	82.24%	85.77%	79.46%	85.34%	82.49%	80.00%	84.51%	73.47%	86.53%	76.60%
80-20	81.12%	83.42%	78.39%	83.94%	80.83%	80.61%	85.29%	73.98%	87.24%	79.23%
85-15	79.21%	82.61%	75.50%	83.22%	78.89%	78.23%	84.30%	69.39%	87.87%	76.12%
90-10	75.00%	81.25%	65.65%	84.54%	72.62%	80.81%	88.46%	70.41%	90.82%	78.41%

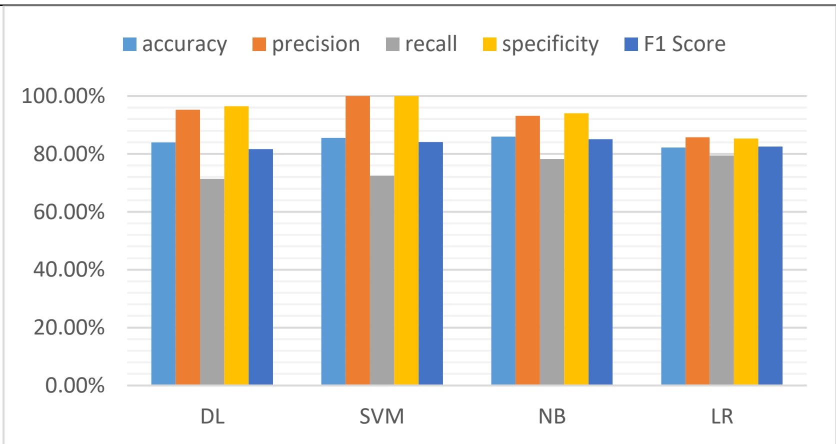

Figure 7. Comparison among of accuracy, precision, recall, specificity and F1 Score

By comparing with previous studies (table 1), this research was able to improve accuracy compared to studies [15], [27], [28], [30]. However, the accuracy achieved is slightly lower compared to study [5], [29].

Osteoporosis has well-known indicators, such as bone density loss, family history, and lifestyle factors, which can be mapped into structured questions. A rule-based chatbot can efficiently guide users through a series of diagnostic questions to identify potential risks without requiring complex AI algorithms [38]. Additionally, its deterministic nature ensures consistent and accurate responses, reducing the risk of misinterpretation. Since osteoporosis screening often relies on basic risk assessment before clinical tests, a rule-based system can act as an effective preliminary tool, referring users to further medical consultation when necessary [39]. Testing was conducted using rule-based chatbots, and the choice of NBC was based on its performance in previous simulations. Based on the set of testing metrics using chatbot, the model demonstrates strong overall performance, with a good balance between identifying positive and negative instances while maintaining high accuracy. In comparison with previous studies (table 1), this research demonstrated improved accuracy over studies [32], [33]. However, the accuracy attained is marginally lower than that reported in studies [31], [34].

Additionally, the study benefits from advanced chatbot features that allow users to independently detect osteoporosis, which other studies do not have. The chatbot includes functionality for handling approximate inputs with terms like "approximately," "about," "around," and the "+/-" symbol for numeric data (age, height, weight), as well as a BMI calculator and information on high-calcium foods, vitamin D intake, and active physical activities.

However, compared to more complex chatbots discussed in [31], [40], [34], the chatbot used in this study is relatively simpler, focusing on rule-based processing for both the order and number of features. While the rule-based chatbot is functional for osteoporosis detection, there are some limitations. Rule-based chatbots are constrained by the predefined rules and logic they operate on, meaning they cannot handle queries or symptoms outside of their programmed scope. This can limit the chatbot's adaptability and its ability to provide personalized responses in more complex or ambiguous medical cases. Additionally, rule-based systems lack learning capabilities, meaning they do not improve over time or from new data, unlike machine learning-based chatbots [41]. As a result, their ability to handle nuanced or evolving medical knowledge is limited, potentially reducing their effectiveness in long-term healthcare applications.

While the chatbot demonstrated effectiveness in a controlled testing environment, it is essential to acknowledge the limitations of this approach regarding its applicability in real-world clinical setting. The controlled environment, while valuable for initial evaluations, does not fully capture the complexities and variabilities of actual patient interactions.

Real-world testing would provide crucial insights into how the chatbot performs with diverse patient populations, accounting for varying levels of health literacy, emotional responses, and the nuances of individual health concerns. Factors such as user experience, accessibility, and the chatbot's ability to manage unexpected queries or provide empathetic responses are critical to its success in a clinical context.

To address these gaps, future studies should prioritize implementing user testing within clinical settings. This would involve collecting feedback from both patients and healthcare professionals to assess the chatbot's performance, usability, and overall impact on patient engagement and health outcomes. Such an approach would not only validate the findings from the controlled environment but also guide necessary refinements to enhance the chatbot's effectiveness and user experience in real-world applications.

5. Conclusion

The dominant features identified using RFE are age, gender, hormonal changes, BMI, calcium intake, vitamin D intake, physical activity, smoking, medications, and prior fractures. In the validation tests for osteoporosis detection, the Naïve Bayes classifier outperformed DL, SVM, and LR models. Additionally, the use of rule-based chatbots, which accommodate flexible numeric inputs, greatly supports users in independently assessing their risk of osteoporosis.

However, this study has several limitations. The most notable is the small sample size, which reduces the statistical power of the findings and limits the ability to detect more nuanced associations between features and osteoporosis risk. Furthermore, the small sample may not adequately represent the broader population, as demographic factors like age, gender, and health behaviors (e.g., calcium intake or physical activity) may vary significantly across different populations. This potential sampling bias impacts the external validity and generalizability of the results.

Despite these limitations, this study contributes valuable insights into osteoporosis detection by demonstrating the effectiveness of the NBC within flexible, rule-based systems. However, addressing the limitations outlined, particularly expanding the sample size, is essential to improve the model's clinical utility and ensure its broader applicability.

 ISSN 2723-6471

Further research is necessary to validate its role in medical practice, with potential for a significant impact on early osteoporosis detection and management.

6. Declarations

6.1. Author Contributions

Conceptualization: E.I.S., R.; Methodology: A.A., W.S.U.; Software: E.I.S., A.A.; Validation: R., W.S.U.; Formal Analysis: E.I.S., A.A.; Investigation: R., W.S.U.; Resources: E.I.S., R.; Data Curation: A.A., W.S.U.; Writing—Original Draft Preparation: E.I.S., A.A.; Writing—Review and Editing: E.I.S., R., W.S.U.; Visualization: A.A., W.S.U.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors extend their gratitude to the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia for funding this research under the Fundamental Research scheme in 2024 (Contract Number: 0609.14/LL5INT/AL.04/2024; 02.3/UTY/KLP/BP/VI/2024).

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. I. Sela and R. Pulungan, "Osteoporosis identification based on the validated trabecular area on digital dental radiographic images," *Procedia Computer Science*, vol. 2019, no. 8, pp. 1-5, 2019, doi: 10.1016/j.procs.2019.08.168.

[2] R. Widyaningrum, E. I. Sela, R. Pulungan, and A. Septiarini, "Automatic segmentation of periapical radiograph using color histogram and machine learning for osteoporosis detection," *International Journal of Dentistry*, vol. 2023, no. 1, pp. 1-9, 2023, doi: 10.1155/2023/6662911.

[3] R. Widyaningrum, S. Lestari, and F. Jie, "Image analysis of periapical radiograph for bone mineral density prediction," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 4, pp. 2083–2090, 2018, doi: 10.11591/ijece.v8i4.pp2083-2090.

[4] E. I. Sela and Sutarman, "Extracting the potential features of digital panoramic radiograph images by combining radio morphometry index, texture analysis, and morphological features," *Journal of Computer Science*, vol. 14, no. 2, pp. 144–152, 2017, doi: 10.3844/jcssp.2018.144.152.

[5] N. Yamamoto, S. Sukegawa, A. Kitamura, R. Goto, T. Noda, K. Nakano, K. Takabatake, H. Kawai, H. Nagatsuka, K. Kawasaki, Y. Furuki, and T. Ozaki, "Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates," *Biomolecules*, vol. 10, no. 11, pp. 1–13, 2020, doi: 10.3390/biom10111534.

[6] A. Aibar-Almazán, A. Voltes-Martínez, Y. Castellote-Caballero, D. F. Afanador-Restrepo, M. del C. Carcelén-Fraile, and E. López-Ruiz, "Current status of the diagnosis and management of osteoporosis," *International Journal of Molecular Sciences*, vol. 23, no. 16, pp. 1–27, 2022, doi: 10.3390/ijms23169465.

[7] C. McPhee, I. O. Aninye, and L. Horan, "Recommendations for improving women's bone health throughout the lifespan," *Journal of Women's Health*, vol. 31, no. 12, pp. 1671–1676, 2022, doi: 10.1089/jwh.2022.0361.

[8] A. J. Singer, A. Sharma, C. Deignan, and L. Borgermans, "Closing the gap in osteoporosis management: The critical role of primary care in bone health," *Current Medical Research and Opinion*, vol. 39, no. 3, pp. 387–398, 2023, doi: 10.1080/03007995.2022.2141483.

[9] S. Laumer, C. Maier, and G. Fabian, "Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis," *27th European Conference on Information Systems (ECIS)*, vol. 2020, no. 6, pp. 1–18, 2020.

[10] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *Journal of Big Data*, vol. 8, no. 1, pp. 1–16, 2021, doi: 10.1186/s40537-021-00413-1.

[11] D.-H. Kim, H.-S. Im, J.-H. Hyeon, and J.-W. Jwa, "Development of the rule-based smart tourism chatbot using Neo4J graph database," *International Journal of Internet, Broadcasting and Communication*, vol. 13, no. 2, pp. 179–186, 2021, doi: 10.7236/IJIBC.2021.13.2.179.

[12] S. A. Thorat and V. Jadhav, "A review on implementation issues of rule-based chatbot systems," *SSRN Electronic Journal*, vol. 2020, no. 5, pp. 1–6, 2020, doi: 10.2139/ssrn.3567047.

[13] R. Franciotti, M. Moharrami, A. Quaranta, M. E. Bizzoca, A. Piattelli, G. Aprile and V. Perrotti, "Use of fractal analysis in dental images for osteoporosis detection: A systematic review and meta-analysis," *Osteoporosis International*, vol. 32, no. 6, pp. 1041–1052, 2021, doi: 10.1007/s00198-021-05852-3.

[14] J. Wang, B. Shu, D. Z. Tang, C. G. Li, X. W. Xie, L. J. Jiang, X. B. Jiang, B. L. Chen, X. C. Lin, X. Wei, X. Y. Leng, Z. Y. Liao, B. L. Li, Y. Zhang, X. J. Cui, Q. Zhang, S. Lu, Q. Shi, and Y. J. Wang, "The prevalence of osteoporosis in China, a community-based cohort study of osteoporosis," *Frontiers in Public Health*, vol. 11, no. Feb 2023, pp. 1–10, 2023, doi: 10.3389/fpubh.2023.1084005.

[15] R. Jang, J. H. Choi, N. Kim, J. S. Chang, P. W. Yoon, and C. H. Kim, "Prediction of osteoporosis from simple hip radiography using deep learning algorithm," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-99549-6.

[16] N. Hong et al., "Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography," *Journal of Bone and Mineral Research*, vol. 38, no. 6, pp. 887–895, 2023, doi: 10.1002/jbmr.4814.

[17] Y. W. Jiang, X. J. Xu, R. Wang, and C. M. Chen, "Radiomics analysis based on lumbar spine CT to detect osteoporosis," *European Radiology*, vol. 32, no. 11, pp. 8019–8026, 2022, doi: 10.1007/s00330-022-08805-4.

[18] D. Qiao, X. Liu, R. Tu, X. Zhang, X. Qian, H. Zhang, J. Jiang, Z. Tian, Y. Wang, X. Dong, Z. Luo, X. Liu, H. Tian, G. Zhang, J. Pan, and C. Wang, "Gender-specific prevalence and influencing factors of osteopenia and osteoporosis in Chinese rural population: The Henan Rural Cohort Study," *BMJ Open*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1136/bmjopen-2018-028593.

[19] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.

[20] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 1–20, 2021, doi: 10.1007/s42979-021-00815-1.

[21] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of support vector machine algorithm in big data background," *Mathematical Problems in Engineering*, vol. 2021, no. 1, pp. 1–10, 2021, doi: 10.1155/2021/5594899.

[22] W. Lian, G. Nie, B. Jia, D. Shi, Q. Fan, and Y. Liang, "An intrusion detection method based on decision tree-recursive feature elimination in ensemble learning," *Mathematical Problems in Engineering*, vol. 2020, no. 1, pp. 1–12, 2020, doi: 10.1155/2020/2835023.

[23] E. Senan, M. H. Al-Adhaileh, F. W. Alsaade, T. H. H. Aldhyani, A. A. Alqarni, N. Alsharif, M. I. Uddin, A. H. Alahmadi, M. E. Jadhav, and M. Y. Alzahrani, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *Journal of Healthcare Engineering*, vol. 2021, no. 1004767, pp. 1–10, 2021.

[24] K. M. Mangano, S. E. Noel, C.-Q. Lai, J. J. Christensen, J. M. Ordovas, B. Dawson-Hughes, K. L. Tucker, and L. D. Parnell, "Diet-derived fruit and vegetable metabolites show sex-specific inverse relationships to osteoporosis status," *Bone*, vol. 144, no. March 2021, pp. 1–9, 2021

[25] R. Sebro and C. De La Garza-Ramos, "Support vector machines are superior to principal components analysis for selecting the optimal bones' CT attenuations for opportunistic screening for osteoporosis using CT scans of the foot or ankle," *Osteoporosis and Sarcopenia*, vol. 8, no. 3, pp. 112–122, 2022.

[26] M. R. Mookiah, "Feasibility of opportunistic osteoporosis screening in routine contrast-enhanced multi detector computed tomography (MDCT) using texture analysis," *Osteoporosis International*, vol. 29, no. 4, pp. 825–835, 2018.

[27] K. Thawnashom, P. Pornsawad, and M. Bunjira, "Machine learning's performance in classifying postmenopausal osteoporosis Thai patients," *Intelligent Medicine*, vol. 7, no. 100099, pp. 1–7, 2023.

[28] V. V. Khanna, "A decision support system for osteoporosis risk prediction using machine learning and explainable artificial intelligence," *Heliyon*, vol. 9, no. 12, pp. e22456-e22468, 2023, doi: 10.1016/j.heliyon.2023.e22456.

[29] Y. Kwon, "Osteoporosis pre-screening using ensemble machine learning in postmenopausal Korean women," *Healthcare*, vol. 10, no. 6, pp. 1–12, 2022, doi: 10.3390/healthcare10061107.

[30] S. Deshpande and J. Warren, "Self-harm detection for mental health chatbots," *Public Health Informatics Proceedings MIE*, vol. 2021, no. 1, pp. 48–52, 2021, doi: 10.3233/SHTI210118.

[31] M. Ghaleb, "Mining the chatbot brain to improve COVID-19 bot response accuracy," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 2619–2638, 2022, doi: 10.32604/cmc.2022.020358.

[32] M. Gao, X. Liu, A. Xu, and R. Akkiraju, "Chatbot or chat-blocker: Predicting chatbot popularity before deployment," in *The 2021 ACM Designing Interactive Systems Conference*, New York, vol. 2021, no. 1, pp. 1458–1469.

[33] S. Chakraborty, "An AI-based medical chatbot model for infectious disease prediction," *IEEE Access*, vol. 10, no. 1, pp. 128469–128483, 2022, doi: 10.1109/ACCESS.2022.3227208.

[34] K. Phasinam, T. Mondal, D. Novaliendry, Y. Cheng-Hong, C. Dutta, and M. Shabaz, "Retracted: Analyzing the performance of machine learning techniques in disease prediction," *Journal of Food Quality*, vol. 2023, no. 9852606, pp. 1–8, 2023.

[35] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Computers and Operations Research*, vol. 135, no. 105456, pp. 1–12, 2021.

[36] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, no. 1, pp. 5198–5219, 2021.

[37] A. Talyshinskii, N. Naik, Z. Hameed, B. M. Juliebø-Jones, and B. K. Somani, "Potential of AI-driven chatbots in urology: Revolutionizing patient care through artificial intelligence," *Current Urology Reports*, vol. 25, no. 1, pp. 9–18, 2024.

[38] C. Quintens et al., "End-users feedback and perceptions associated with the implementation of a clinical-rule based Check of Medication Appropriateness service," *BMC Medical Informatics and Decision Making*, vol. 22, no. 177, pp. 1–10, 2022.

[39] S. Roca, J. Sancho, J. García, and Á. Alesanco, “Microservice chatbot architecture for chronic patient support,” *J. Biomed. Inform.*, vol. 102, no. May 2019, pp. 1–9, 2020, doi: 10.1016/j.jbi.2019.103305

[40] C. Papakostas, C. Troussas, A. Krouska, and S. Cleo, "A rule-based chatbot offering personalized guidance in computer programming education," *Lecture Notes in Computer Science*, vol. 14799, no. 1, pp. 1–15, 2024.

[41] S. Zhang and J. Song, "A chatbot-based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model," *Nature Portfolio*, vol. 14, no. 17118, pp. 1–10, 2024.