International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I1P114

Volume 73 Issue 1, 166-176, January 2025

Original Article

Evolutionary Feature Selection to Classify Elderly Diseases from Dietary and Exercise Habits and Emotions

Nitaya Buntao[1], Rada Somkhuean[2], Wongpanya S. Nuankaew[3], Pratya Nuankaew[4]

[1]*Department of Applied Statistics, Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand.*
[2]*Department of Mathematics, Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Chiang Mai, Thailand.*
[3]*Department of Computer Science, School of Information and Communication Technology, University of Phayao, Phayao, Thailand.*
[4]*Department of Digital Business, School of Information and Communication Technology, University of Phayao, Phayao, Thailand.*

[4]*Corresponding Author : pratya.nu@up.ac.th*

Received: 02 August 2024 Revised: 11 December 2024 Accepted: 17 December 2024 Published: 31 January 2025

Abstract *- The research investigates the effectiveness of various feature selection methods in enhancing disease classification models for elderly populations based on dietary habits, physical activity, and emotional well-being. It is conducted in Maha Sarakham Province, Thailand, and addresses critical healthcare challenges specific to this demographic. Traditional greedy algorithms (Forward Selection, Backward Elimination) are contrasted with metaheuristic approaches like evolutionary feature selection, evaluating their impact on accuracy and model robustness across classification algorithms (Deep Learning with H2O, Naïve Bayes, Gradient Boosted Trees, KNN, Decision Trees, Generalized Linear Models). Results show that evolutionary feature selection consistently outperforms traditional methods, achieving an average accuracy of 79.69% with Logistic Regression and Generalized Linear Models and demonstrating a superior balance between precision and recall. Deep Learning with H2O performs strongly across all methods, while Naïve Bayes benefits from Backward Elimination. The findings highlight the potential of evolutionary feature selection to enhance disease classification accuracy and model reliability, emphasizing the need for personalized healthcare strategies tailored to individual profiles in older adults.*

Keywords *- Evolutionary feature selection, Meta heuristic approaches, Personalized healthcare strategies, Elderly diseases, Healthcare challenges, Classification algorithms.*

1. Introduction

The world's senior population is booming, with the World Health Organization (WHO) predicting a doubling by 2050 [1]. This translates to a rise in chronic illnesses among older adults, which affect roughly 80% of this demographic [2]. Chronic conditions like cardiovascular diseases, diabetes, chronic respiratory disorders, and dementia take a toll on both quality of life and healthcare systems, and It is evident that innovative approaches to elder care and disease management are needed. The key to tackling this challenge lies in understanding how daily habits. People already know regular diet, exercise, and emotional well-being impact health. Research shows regular exercise can significantly reduce chronic disease risk, improve mental health, and boost physical function in older adults [3]). Similarly, studies highlight the positive effects of balanced nutrition and exercise on managing conditions like hypertension [4]. Beyond physical health, emotional well-being plays a crucial role. Research suggests managing negative emotions might be more effective than simply boosting positive ones to reduce inflammation in older adults [5]. It opens doors for exploring how emotional regulation can contribute to overall health. Previous research has explored the application of machine learning (ML) algorithms for early diagnosis of chronic diseases, emphasizing their potential to improve patient outcomes and treatment strategies. Feature selection is critical in developing accurate machine-learning models [6, 7]. The main feature selection methods include filtering, wrapping, and Embedding [8]. Advanced techniques like Genetic Algorithms (GAs) optimize feature subsets for complex datasets. Feature selection enhances model performance, reduces computational costs, and improves interpretability. The choice of method depends on the dataset size, interpretability requirements, and computational capacity [9]. This study identifies crucial features for accurate, efficient machine learning models in elderly disease classification. It

aims to enable proactive health management, potentially enhancing the aging population's quality of life. Accurately predicting diseases like mild cognitive impairment (MCI) in older adults is crucial. This research tackles this challenge by focusing on feature selection, a critical step in building robust machine-learning models. Finding the most important data: Forward Selection, Backward Elimination, Stepwise Selection, and Features Finally Used. Evaluate models directly to identify the most impactful features, leading to the best predictions [10]. Effective feature selection, like focusing on diet and exercise, helps build efficient and accurate models for predicting diseases in older adults. Studies show the positive impact of healthy habits. Research by Nitschke et al. (2022) analysed 82 studies and found that interventions promoting nutrition and physical activity improved weight, blood pressure, and blood sugar - all factors crucial for reducing chronic disease risk [11].

It highlights the importance of lifestyle interventions for overall health. For doctors to use these models effectively, clear explanations are essential. Abbas Saad Alatrany et al. (2024) proposed an approach for Alzheimer's disease (AD) classification that achieves high accuracy and provides clear explanations. It allows doctors to understand better the factors influencing the diagnosis [12]. In conclusion, effective feature selection and explainable machine-learning approaches are essential for improving disease prediction and management. These methods and lifestyle interventions play a critical role in promoting overall health and preventing chronic diseases. Advanced computational techniques, particularly evolutionary feature selection, effectively analyze large datasets to classify diseases among the elderly. These methods use diet, physical activity, and emotional state data to predict health outcomes accurately. Khanna et al. (2024) introduced a computer-aided diagnosis system for breast cancer classification using Teaching Learning-Based Optimization and Elephant Herding Optimization.

These methods improved classification accuracy and reduced unnecessary features [13]. Rashid et al. (2022) developed an AI-based method for chronic disease prediction, integrating Artificial Neural Networks with Particle Swarm Optimization. This approach focused on diseases like breast cancer, diabetes, and heart attack, outperforming traditional methods [14]. De Lacy et al. (2022) explored integrated evolutionary learning for complex medical datasets, automating features, and hyperparameter selection. These techniques aim to enhance the accuracy, reliability, and interpretability of disease diagnosis models for the elderly, potentially revolutionizing healthcare strategies for this growing population [15]. This research explores machine learning for identifying disease risks in older adults from Maha Sarakham, Thailand. It focuses on feature selection to pick crucial data (diet, activity, emotions) for accurate models. The research compares greedy algorithms (e.g., Forward Selection, Backward Elimination) with genetic algorithms (Evolutionary Method) to find the most effective approach for disease classification, aiming to improve early detection and health outcomes.

- Identify key factors contributing to disease prediction across different approaches for the elderly.
- Conduct a comparative analysis of these algorithms in elderly disease classification.
- Explore method synergies to enhance model robustness and accuracy.

This research contributes to geriatric health informatics by developing interpretable models. By explaining their reasoning, these models can inform personalized healthcare strategies in Thailand and globally.

2. Literature Review

Disease classification in elderly populations is vital for enhancing healthcare outcomes. Traditional feature selection methods, such as greedy algorithms (e.g., forward selection, backward elimination), sequentially select features based on individual contributions to model performance. However, these methods often lack feature interactions and struggle with large datasets. Advanced feature selection methods, like evolutionary algorithms, offer a more sophisticated approach. Inspired by natural selection, they explore and select features by considering their interactions and impact on model performance. This approach improves model accuracy by reducing dimensionality and identifying complex feature interactions. Effective feature selection is essential for developing accurate and interpretable machine learning models. Identifying key predictors allows researchers to tailor healthcare strategies for the elderly. The literature reveals a growing trend of using machine learning and Artificial Intelligence for disease classification and prediction in elderly populations, particularly for chronic conditions like Parkinson's, diabetes, and heart disease. Future directions could include integrating domain knowledge with evolutionary algorithms or developing more advanced feature selection techniques to enhance model performance further and advance personalized medicine for the elderly.

2.1. Disease Classification in Elderly Populations

Recent studies highlight a shift towards advanced computational methods for disease classification in geriatric medicine. Khera and Kumar (2020) proposed an ensemble learning classifier with optimal feature selection for Parkinson's disease, showcasing sophisticated algorithmic approaches in geriatric medicine [16]. Similarly, Qin et al. (2022) developed machine learning models for predicting diabetes based on lifestyle factors, emphasizing the importance of data-driven methods in managing chronic diseases in older adults [17]. Ali et al. (2023) also investigated Parkinson's disease detection using filter feature selection and genetic algorithms combined with ensemble learning, highlighting advanced computational methods in neurological

disorder diagnosis [18]. Further, Chawla et al. (2024) and Bhakar et al. (2024) focused on Parkinson's disease classification, employing nature-inspired feature selection methods and hybrid models, respectively. These studies demonstrate ongoing refinement in classification techniques for age-related neurological disorders [19, 20]. Collectively, these studies indicate a shift towards personalized and precise disease classification methods for elderly patients, leveraging machine learning to enhance diagnostic accuracy and inform tailored treatment strategies.

2.2. Existing Approaches to Feature Selection

2.2.1. Traditional Methods (Greedy Algorithms)

Feature selection is vital in developing predictive models for medical conditions like Chronic Kidney Disease (CKD) and Mild Cognitive Impairment (MCI). Traditional methods, including Filter, Wrapper, and Embedded approaches, provide foundational techniques in this area. These methods, such as forward selection, backward elimination, and stepwise regression, often serve as baseline comparisons in research studies. Recent investigations have demonstrated the significant impact of feature selection on model accuracy. For example, Zeynu and Patil (2018) showed that feature selection techniques substantially improved the precision of CKD prediction models. Their research used both Filter and Wrapper methods to refine the dataset and identify key attributes. Additionally, they implemented an ensemble model integrating multiple classifiers through a voting mechanism, enhancing prediction performance [21].

Similarly, Lim S-J et al. (2021) explored feature selection in predicting MCI using medical records. Their approach incorporated both Filter and Wrapper methods. The Filter method assessed individual features based on relevance, while the Wrapper method employed recursive elimination to identify optimal feature subsets. These strategies aimed to boost prediction accuracy by focusing on essential attributes and reducing dimensionality. The study also compared various classifiers to evaluate the impact of feature selection on model performance [10]. Additionally, Purwaningsih (2022) utilized forward selection to predict CKD, a technique that iteratively adds features to a Support Vector Machine (SVM) model based on their impact on performance. This approach seeks to identify the most relevant features for CKD detection, thereby enhancing the SVM's effectiveness. Despite its benefits, forward selection has limitations, including the potential for reduced generalizability due to the small dataset size and the possibility of overlooking important feature interactions. Broader feature selection methods or additional validation techniques could address these issues, potentially improving model robustness and applicability across various datasets [22]. More recently, K Hema et al. (2024) investigated feature selection techniques for early CKD prediction, employing Filter, Wrapper, and Embedded methods. Their study demonstrated that advanced feature selection methods improved prediction accuracy. The Filter method evaluated individual features' relevance, while the Wrapper method refined feature subsets through iterative approaches. Embedded methods optimized feature selection during model training. However, limitations included a lack of exploration of diverse techniques and potential overfitting due to dataset constraints. Future research should address these limitations by expanding feature selection techniques and testing on more varied datasets to enhance model robustness and generalizability [23]. In conclusion, while traditional feature selection methods have shown promise, there is significant room for advancement. By addressing current limitations and exploring more advanced techniques, future studies can contribute to developing even more accurate and reliable predictive models in healthcare.

2.2.2. Advanced Methods (Evolutionary Algorithms)

Recent research demonstrates the growing prominence of advanced, often nature-inspired or evolutionary approaches: The prominence of advanced feature selection algorithms in medical diagnostics is evident in recent research. Abdollahi and Nouri-Moghaddam (2021) evaluated a hybrid Stacked-Genetic approach for heart disease diagnosis, illustrating the integration of evolutionary algorithms in medical feature selection [24]. This trend highlights the effectiveness of evolutionary and nature-inspired algorithms in medical diagnosis and prediction tasks, often outperforming traditional greedy algorithms in accuracy and robustness for elderly disease classification.

Moreover, de Lacy et al. (2022) introduced an integrated evolutionary learning approach that simultaneously optimizes both feature selection and model parameters, showcasing a sophisticated method for medical diagnostics [15]. Ali et al. (2023) combined genetic algorithms with filter feature selection, demonstrating a hybrid approach that blends traditional and evolutionary methods [18]. Similarly, Chawla et al. (2023) employed nature-inspired feature selection techniques, moving towards bio-inspired optimization in healthcare data analysis [19]. Bhakar et al. (2024) also proposed a hybrid model incorporating random classification and feature selection, indicating a trend of combining multiple advanced techniques for enhanced performance [20]. Collectively, these studies reflect a shift towards more personalized and precise disease classification methods for elderly patients, leveraging machine learning to improve diagnostic accuracy and inform tailored treatment strategies.

2.3. Studies on the Impact of Diet, Exercise, and Emotional Health on Disease Outcomes in the Elderly

While most studies focus on computational methods, some address the impact of lifestyle factors on health outcomes: Studies on the Impact of Diet, Exercise, and Emotional Health on Disease Outcomes in the Elderly: While most of the provided papers focus on computational methods for disease classification, some address the impact of lifestyle factors on health outcomes in elderly populations. For

instance, Qin et al. (2022) developed machine learning models for diabetes prediction based on lifestyle types, implicitly considering factors such as diet and exercise in their analysis [17]. Additionally, Rashid et al. (2022) proposed an augmented artificial intelligence approach for chronic disease prediction, likely incorporating lifestyle factors as part of its predictive model [14].

Furthermore, while focusing on breast cancer, Khanna et al. (2024) developed an enhanced approach for chronic human disease prediction that could potentially be applied to lifestyle-related conditions in the elderly [13]. While traditional feature selection methods have proven valuable, several areas warrant further exploration. First, studies on more diverse elderly populations are needed to ensure the generalizability of findings. Second, incorporating environmental and lifestyle factors, such as diet, exercise, and sleep quality, could offer a more holistic understanding of disease risk in this population. Feature selection techniques have already shown promise in identifying the most impactful factors within these domains (e.g., Khanna et al., 2024). Finally, exploring more advanced feature selection techniques beyond traditional methods holds the potential to further improve model accuracy and robustness.

3. Methodology

3.1. Feature Selection Algorithm

In the realm of machine learning, feature selection stands as a cornerstone in model development, significantly enhancing efficiency and mitigating complexity. The research methodology often incorporates three principal approaches: Evolutionary Algorithm (EA), Forward Selection (FS), and Backward Elimination (BE). Evolutionary feature selection improves classification by identifying multiple optimal feature subsets through complex interactions using heuristic search methods. Techniques include multimodal optimization, differential evolution, duplication analysis, niching-based, binary differential evolution, and feature clustering-assisted selection. These methods select smaller feature subsets while maintaining accuracy, generating diverse non-dominated solutions, and reducing redundancy. Solutions with high diversity scores enhance population diversity. This approach excels in navigating intricate search spaces, making it effective for handling complex datasets [25, 26]. Forward feature selection begins with an empty feature set and adds features incrementally based on their contribution to model performance. Initially, it selects the feature that improves performance the most. The algorithm then evaluates combinations of the selected and remaining features, adding the feature with the highest performance boost. This process continues until a stopping criterion is met. It aims to maximize classification accuracy or minimize error rates while being computationally efficient and suitable for large datasets. However, its greedy approach may not always yield the optimal subset, and its success depends on the chosen selection criteria [27]. In contrast, Backward Elimination is a feature selection method that improves predictive models by removing the least significant features based on their statistical impact. By retaining only the most relevant variables, this technique enhances model accuracy and generalization. It simplifies the model, making it easier to interpret and less resource-intensive. Additionally, it speeds up the training process by reducing the number of features, leading to greater efficiency [28]. While these methodologies offer versatility across various data types, their efficacy is inherently tied to the specific research context. Astute researchers must carefully weigh factors such as dataset dimensions, problem intricacy, and available computational resources when selecting the most appropriate feature selection technique for their unique challenges.

3.2. Classification Algorithm

In academic research, various classification algorithms are employed to analyze and interpret complex datasets. These algorithms range from simple, intuitive methods to sophisticated machine-learning techniques, each with its own strengths and limitations. Deep Learning algorithms, particularly those implemented using platforms like H2O, represent the cutting edge of machine learning. These algorithms utilize multi-layered neural networks to extract features and learn from data, making them exceptionally adept at handling complex, high-dimensional datasets. They excel in tasks such as image classification and natural language processing. However, their power comes at a cost: they typically require large datasets, involve time-consuming training processes, and can produce results that are challenging to interpret [29].

On the other end of the spectrum, algorithms like Naïve Bayes operate on simpler principles. Naïve Bayes employs Bayes' probability theorem, assuming independence between features. This approach is user-friendly, computationally efficient, and effective for small to medium-sized datasets, making it particularly useful for tasks like text classification and spam detection. However, its underlying assumption of feature independence may not always hold in real-world scenarios [30]. Gradient Boosted Trees offer a middle ground, combining multiple decision trees to create powerful predictive models. This method effectively manages complex and imbalanced datasets, making it suitable for data with non-linear relationships and numerous features.

However, careful parameter tuning is required to avoid overfitting, which can involve lengthy training periods [31]. The K-Nearest Neighbors (KNN) algorithm provides an intuitive approach to classification, basing its decisions on the K nearest data points in the training set. While it's easy to implement and makes no assumptions about data distribution, its performance can degrade with high-dimensional data, and prediction times increase for large datasets [32]. Decision Trees offer a highly interpretable model, constructing a tree-like structure where each node represents a feature-based

decision. This approach is particularly valuable when model explainability is crucial. However, Decision Trees are prone to overfitting, especially when allowed to grow too deep [33].

For data that follows specific probability distributions, Generalized Linear Models (GLMs) extend the concepts of linear regression beyond normal distributions. GLMs are flexible and capable of elucidating variable relationships, but they may struggle with highly complex, non-linear relationships and require statistical expertise for proper interpretation [34]. Logistic regression is a tool for binary classification, like predicting heart disease. It evaluates how risk factors (such as high cholesterol and smoking) relate to the likelihood of developing cardiovascular disease.

The model computes probabilities based on input features and classifies individuals accordingly. Stored with the Pickle library for convenient deployment and reuse, it uses a logistic function to convert features into probabilities for binary predictions [35]. When selecting an appropriate algorithm for classification tasks, it is crucial to consider various factors, including the characteristics and size of the dataset, the complexity of the problem at hand, requirements for result interpretation, and available computational resources.

Experimenting with multiple algorithms and comparing their performance using metrics such as Accuracy, F1-score, or Area Under the ROC Curve (AUC-ROC) is often beneficial. This empirical approach allows the identification of the most suitable model for specific research endeavors, balancing predictive power with interpretability and computational efficiency.

3.3. Preprocessing

3.3.1. Data Collection

This research collected data from 215 elderly individuals aged 60 and above residing in the Kang Leung Chan Sub-district, Mueang District of Maha Sarakham Province. These participants were selected from a total population of 1,505 elderly individuals in Maha Sarakham Province, Thailand, between 2021 and 2022. The data collection tool was a researcher-adapted questionnaire consisting of two sections:

Section 1: General Information Questionnaire

This section includes questions on gender, age, weight, height, marital status, educational level, occupation, income, source of income, marital status, living conditions, household status, caregiver, history of alcohol consumption, smoking history, and chronic diseases. Respondents are asked to fill in or select the information that corresponds to their own.

Section 2: Questionnaire on Eating Habits, Exercise, and Mood

This section features 19 questions where respondents select answers by marking the appropriate box. It uses a rating scale to classify behaviors into three levels: regular (5 - 7 days/week), occasional (1 - 4 days/week), and never. The assessed behaviors include smoking, sleeping, and drinking water, as outlined in Table 1. The researchers conducted the data collection process by explaining the purpose of the data collection and describing the nature of the questionnaire, including how to respond. The researchers personally gathered the data through interviews, allowing participants to complete the questionnaire themselves. The researchers then verified the accuracy and completeness of the questionnaires, recorded the data, and documented the process with photographs as evidence.

Table 1. Questionnaire on Eating Habits, Exercise, and Mood

Question	Feature name	Choice
Do you smoke?	Smoking	Yes No
How many hours did you sleep per night on average in the past week?	Sleep_per_night	Less than 5 hours/night 5 – 6 hours/night 7 – 8 hours/night
How often do you drink at least 8 glasses of water per day in a week?	Drink_water_per_day	1 – 3 days/a week 4 – 6 days/a week 7 days /a week
Eating habits, exercise, and mood behaviors		
Do you consume a balanced diet consisting of all five food groups (meat-dairy-eggs, grains, vegetables, fruits, and oils)?	Feature_Q1	Regular (5-7 days/week) Occasional (1-4 days/week) Never
Do you have breakfast as your main meal?	Feature_Q2	
Do you eat at least six servings of vegetables per day?	Feature_Q3	
Do you eat 4-5 servings of fruit per day (one serving equals 6-8 bites)?	Feature_Q4	
Do you eat fish at least once a day?	Feature_Q5	
Do you eat lean meat 2-3 times per week?	Feature_Q6	
Do you drink plain milk, low-fat milk, skim milk, or unsweetened soy milk with black sesame once or twice a day?	Feature_Q7	
Do you eat dinner at least 4 hours before bedtime?	Feature_Q8	

Do you consume foods that are boiled, steamed, blanched, baked, or grilled?	Feature_Q9	
Do you avoid high-fat foods?	Feature_Q10	
Do you avoid drinks, desserts, and snacks high in flour and sugar or very sweet?	Feature_Q11	
Do you eat bland food?	Feature_Q12	
Do you choose to drink water instead of soda or sweetened beverages?	Feature_Q13	
Do you avoid alcoholic beverages?	Feature_Q14	
Do you maintain a good mood and avoid stress?	Feature_Q15	
Do you sleep at least 7-8 hours per night?	Feature_Q16	
Do you exercise 5 days a week or 5 times a week?	Feature_Q17	
Do you exercise for at least 30 minutes a day?	Feature_Q18	
During exercise, do you breathe faster than usual and break a sweat?	Feature_Q19	

Table 2. Data Preparation

Category	Male	Female
Total Number	82	129
Average Age (years)	69.39	68.62
Average weight (kg)	56.16	57.21
Average height (cm)	160.95	158.36
Has Chronic Illness	43	83
Has More Than 1 Chronic Illness	39	32
No Chronic Illness	18	46
Exercises Regularly	62	112
Smokes	12	9
Sleeps More Than 5 Hours/Day	74	121
Drinks At Least 8 Glasses of Water/Day/Week	81	128

3.3.2. Data Preparation

After collecting data from the questionnaire, researchers proceeded with preparing the data for use in model building, which involved the following steps:

- Data Cleaning:
  Addressing errors and inconsistencies like missing values, duplicates, and outliers.
- Data Transformation:
  Converting data into a suitable format for analysis, including normalization, scaling, encoding categorical variables, and aggregating data.
- Data Integration:
  Consolidating data from different sources into a cohesive dataset, maintaining consistency and integrity.
- Data Reduction:
  Simplifying the dataset by selecting relevant features, aggregating data, and removing redundant or irrelevant information.
- Data Validation:
  Ensuring data accuracy and quality through consistency checks and verification against established benchmarks.
- Data Formatting:
  Structuring data for analysis or modelling, organizing it into tables with appropriate headers and ensuring consistent data types.
- Data Splitting:
  Dividing data into training and testing sets, 70:30 was used to evaluate model and algorithm performance.

These steps ensure the data is accurate, consistent, and ready for analysis, leading to more reliable and meaningful results. The 22-question questionnaire on eating habits, exercise, and mood was used for data modelling, with chronic diseases (Yes/No) as the class label derived from the general information questionnaire. Following data preparation, the dataset comprised 211 elderly individuals, as detailed in Table 2.

3.4. Modelling

After Data Preparation, the modelling process involve comparing feature selection methods: Evolutionary Algorithm (EA), Forward Selection (FS), and Backward Elimination (BE). These assess which variables best enhance model performance. Additionally, various classification algorithms are compared: Deep Learning (H2O), Naïve Bayes, Gradient Boosted Trees, K-Nearest Neighbours (KNN), Decision Trees, Generalized Linear Models, and Logistic Regression. Evaluation criteria include accuracy, precision, recall, F1-score, and computational efficiency to identify the optimal approach for the dataset's needs.

3.4.1. First Objective

Conduct a comparative analysis of greedy algorithms (e.g., Forward Selection, Backward Elimination) versus metaheuristic algorithms (e.g., evolutionary methods) to enhance disease classification accuracy for the elderly. Identify key factors by analysing the overlap and uniqueness of selected variables and assessing the impact of dietary

habits, physical activity, and emotional well-being on disease prediction.

3.4.2. Second Objective

Conducting a comparative analysis of classification models is essential to evaluate the performance of greedy feature selection algorithms versus evolutionary algorithms. This study employs various models, including Deep Learning (H2O), Naïve Bayes, Gradient Boosted Trees, K-Nearest Neighbours (KNN), Decision Trees, Generalized Linear Models, and Logistic Regression, to improve disease classification accuracy for elderly populations. Greedy feature selection algorithms, such as Forward Selection and Backward Elimination, incrementally add or remove features based on immediate performance impact. These methods are efficient but may overlook complex feature interactions. In contrast, evolutionary algorithms use metaheuristic techniques inspired by natural evolution, such as mutation, crossover, and selection, to explore a broader search space. These methods handle complex, high-dimensional datasets effectively, uncovering intricate feature interactions. The analysis aims to identify the most effective feature selection method for improving disease prediction accuracy in elderly populations. It considers dietary habits, physical activity levels, and emotional well-being, comprehensively evaluating factors contributing to disease outcomes. This study offers insights into the strengths and limitations of greedy and evolutionary feature selection methods, guiding the choice of techniques to enhance disease classification accuracy and leading to better health outcomes and targeted interventions. The modeling process in this research is outlined in Algorithm 1.

Algorithm1: Framework process

Input: Training set, Testing set

1. Read the Training Set
2. Define the range of training data (i to j) and attributes (m to n)
3. Define classifiers c (c_1, c_2, ..., c_k)
4. Define attribute weights w (w_1, w_2, ..., w_z)
5. Compute attribute weights w_1 to w_z using Forward Selection, Backward Elimination, and Evolutionary Methods
6. Rank attributes by weight for each feature selection method
7. Define rankings r (r_1, r_2, ..., r_z) from highest to lowest weight for each method
8. Select attributes from ranked list r (1 to z) for the best classification using c_1 to c_k
9. Build classification models using the selected attributes from each feature selection method
10. Read the Testing Set
11. Evaluate the models

Output:

1. Attribute weight values
2. Accuracy, Precision, Recall, and F1 Score values

3.5. Evaluation

The evaluation employs statistical methods to assess the significance of differences in classification performance between greedy and evolutionary algorithms. Model quality was assessed using a 70:30 split of training and testing sets. Efficiency was measured using the following metrics: accuracy, precision, recall, and F1 score, as defined by the equations below [23]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1} = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (4)$$

In these equations, TP, FP, TN, and FN refer to true positive, false positive, true negative, and false negative counts, respectively.

4. Results and Discussion

4.1. Research Results

This study explores feature selection methods for classifying diseases in elderly populations, emphasizing dietary habits, exercise, and emotional well-being as predictors. It compares Evolutionary Algorithms (EA), Forward Selection (FS), and Backward Elimination (BE) to enhance predictive accuracy and guide targeted healthcare interventions for the second objective, as detailed in Table 3. For the first objective, the research focuses on features with a weight value greater than 0.05, as listed below. The Evolutionary Algorithm identified nine key features: Smoking, Feature_Q1, Feature_Q2, Feature_Q3, Feature_Q10, Feature_Q11, Feature_Q13, Feature_Q15, Feature_Q17, and Feature_Q18. This focus underscores the significance of dietary habits and lifestyle in maintaining overall well-being. For dietary habits, the algorithm highlighted essential features such as consuming a balanced diet (Feature_Q1), eating breakfast regularly (Feature_Q2), consuming at least six servings of vegetables daily (Feature_Q3), avoiding high-fat foods (Feature_Q10), and

eating bland food (Feature_Q11). These aspects are crucial for health promotion and disease prevention. In terms of lifestyle, the selected features include smoking, choosing water over sugary drinks (Feature_Q13), getting adequate sleep (Feature_Q15), and engaging in regular exercise (Feature_Q17 and Feature_Q18). These behaviors are vital components of a healthy lifestyle and significantly impact overall health.

The Forward Algorithm selected thirteen key features: Smoking, Sleep_per_night, Drink_water_per_day, Feature_Q1, Feature_Q2, Feature_Q3, Feature_Q4, Feature_Q10, Feature_Q12, Feature_Q11, Feature_Q13, Feature_Q19, and Feature_Q9. This algorithm identifies essential features for evaluating health-related behaviors, focusing on smoking, sleep, hydration, diet, and exercise. It assesses smoking habits, average sleep duration (Sleep_per_night), and daily water intake (Drink_water_per_day), which are crucial for hydration. Dietary aspects include whether the individual consumes a balanced diet (Feature_Q1), eats breakfast regularly (Feature_Q2), and their daily intake of vegetables (Feature_Q3) and fruits (Feature_Q4). The algorithm also evaluates avoidance of high-fat foods (Feature_Q10), preference for bland food (Feature_Q12), avoidance of sugary foods (Feature_Q11), and choosing water over sugary drinks (Feature_Q13). Additionally, it measures exercise intensity (Feature_Q19) and preference for boiled, steamed, or grilled foods (Feature_Q9).

These features provide a comprehensive view of lifestyle factors impacting overall health and well-being, incorporating specific measures such as average sleep hours, daily water consumption, and dietary and exercise-related factors. The Backward Elimination algorithm selected thirteen features: Smoking, Sleep_per_night, Feature_Q2, Feature_Q3, Feature_Q4, Feature_Q7, Feature_Q9, Feature_Q10, Feature_Q11, Feature_Q12, Feature_Q13, Feature_Q14, and Feature_Q16. It has identified key features for assessing health-related behaviors. These include smoking (whether the individual smokes) and sleep_per_night (average hours of sleep). Dietary habits are evaluated through breakfast frequency (Feature_Q2), daily vegetable intake (Feature_Q3), and daily fruit intake (Feature_Q4). Additional factors include milk or soymilk consumption (Feature_Q7), preference for boiled, steamed, or grilled foods (Feature_Q9), and avoidance of high-fat foods (Feature_Q10), sugary foods (Feature_Q11), and alcohol (Feature_Q14) indicates a preference for bland food (Feature_Q12), measures choosing water over sugary drinks (Feature_Q13) and checks if the individual gets at least 7-8 hours of sleep (Feature_Q16).

The Backward Elimination algorithm introduced unique elements like milk or soymilk consumption and alcohol avoidance, which were not featured in the other algorithms, offering additional insights into health-related behaviors. Based on a comprehensive analysis of various algorithms and feature selection methods, evolutionary feature selection consistently achieved the highest accuracy, notably with Logistic Regression and Generalized Linear Models averaging 79.69%. Precision and recall metrics varied across methods, with evolutionary approaches demonstrating superior balance compared to Forward and Backward elimination techniques.

F1 scores, reflecting the harmonic mean of precision and recall, also favored evolutionary methods across diverse algorithms. Deep learning using H2O showed consistently strong performance across all feature selection methods, maintaining high accuracy, precision, recall, and F1 scores with minimal variation. Naïve Bayes performed well in precision and recall, especially enhanced by Backward Elimination. Gradient Boosted Trees, KNN, Decision Trees, and Generalized Linear Models exhibited mixed performance across different feature selection techniques, with evolutionary methods generally providing more stable outcomes.

Overall, Evolutionary Feature Selection emerged as the preferred method due to its superior performance in accuracy, precision, recall, and F1 scores across various classification algorithms. This underscores its potential for optimizing disease classification models in elderly populations based on dietary habits, physical activity, and emotional well-being. Leveraging evolutionary methods, particularly with Logistic Regression and Deep Learning using H2O, is recommended for enhancing model robustness and predictive accuracy in healthcare applications.

Third Objective: Explore synergies between methods to enhance model robustness and accuracy. To improve model performance, especially for elderly populations, integrating Evolutionary Feature Selection with algorithms like Logistic Regression and Deep Learning is proposed.

Table 3. Research Results

	Accuracy	**Precision**		**Recall**		**F1-Score**	
		Yes	**No**	**Yes**	**No**	**Yes**	**No**
Forward Selection							
Deep Learning algorithm using H2O	75.00	72.00	85.71	94.74	46.15	81.82	60.00
Naïve Bayes	71.88	69.23	83.33	94.74	38.46	80.00	52.63
Gradient Boosted Trees	70.31	75.68	62.96	73.68	65.38	74.67	64.15

KNN	70.31	70.17	65.22	78.95	57.69	75.95	61.22
Decision Tree	70.31	68.63	76.92	92.11	38.46	78.65	51.28
Generalized Linear Model	71.88	69.23	83.33	94.74	38.46	80.00	52.63
Logistic Regression	73.44	78.38	66.67	76.23	69.23	77.34	67.92
Backward Elimination							
Deep Learning algorithm using H2O	76.56	81.08	70.37	78.95	73.08	80.00	71.70
Naïve Bayes	75.00	84.38	65.62	71.05	80.77	77.14	72.41
Gradient Boosted Trees	71.88	73.81	68.18	81.58	57.69	77.75	62.49
KNN	71.88	70.83	75.00	89.47	46.15	79.07	57.14
Decision Tree	71.88	72.73	70.00	84.21	53.85	78.05	60.87
Generalized Linear Model	75.00	80.56	67.86	76.32	73.08	78.37	70.37
Logistic Regression	76.56	84.85	67.74	73.68	80.77	78.87	73.68
Evolutionary							
Deep Learning algorithm using H2O	78.12	81.58	73.08	81.58	73.08	81.59	73.08
Naïve Bayes	76.56	82.86	68.97	76.32	76.92	79.45	72.72
Gradient Boosted Trees	73.44	72.34	76.47	89.47	50.00	80.00	60.47
KNN	73.44	76.92	68.00	78.95	65.38	77.92	66.67
Decision Tree	78.12	83.33	71.43	78.95	76.92	81.08	74.07
Generalized Linear Model	79.69	80.49	78.26	86.84	69.23	83.54	73.47
Logistic Regression	79.69	83.78	74.07	81.58	76.92	82.67	75.47

This combined approach leverages the strengths of each method to optimize disease classification models by prioritizing factors such as diet, physical activity, and emotional well-being.

4.2. Discussion

Investigating feature selection methods for disease classification among elderly populations based on dietary habits, exercise routines, and emotional well-being provides crucial insights for advancing healthcare practices. Quantitative analysis highlights evolutionary approaches as particularly effective, with Logistic Regression and Generalized Linear Models achieving notable average accuracies of 79.69%. Evolutionary methods excel in balancing precision and recall metrics compared to traditional Forward and Backward elimination methods, underscoring their superiority. These methods consistently identify predictive factors like dietary habits, exercise routines, and emotional well-being indicators, offering nuanced insights into health outcomes among older adults.

They outperform traditional approaches across various classification algorithms by managing complex feature interactions, thereby enhancing model robustness and predictive accuracy. Conversely, traditional methods often struggle to maintain this balance, potentially overlooking critical relationships between dietary, exercise, and emotional variables. Understanding these synergistic relationships is pivotal for effective disease classification in elderly populations. Evolutionary feature selection effectively captures these dynamics, revealing how specific dietary patterns and exercise frequencies influence both emotional well-being and physical health outcomes. This comprehensive understanding informs tailored healthcare interventions integrating dietary modifications, personalized exercise regimens, and emotional support strategies to improve disease prevention and management among older adults.

5. Conclusion

In conclusion, this study rigorously evaluates various feature selection methods to enhance disease classification models for elderly populations, specifically focusing on dietary, exercise, and emotional factors. Evolutionary Algorithms (EA) are highlighted for consistently achieving superior predictive accuracy, precision, recall, and balanced F1 scores across diverse algorithms, effectively identifying critical predictive features and revealing nuanced relationships between lifestyle factors and health outcomes among older adults.

In contrast, traditional methods like Forward Selection (FS) and Backward Elimination (BE) show variable performance, often grappling with precision-recall trade-offs and occasionally missing subtle yet significant feature interactions. While each method offers valuable insights into feature relevance, Evolutionary Algorithms emerge as the optimal choice for enhancing model robustness and accuracy in complex healthcare scenarios. Moving forward, further research should extend algorithm comparisons to include longitudinal studies and considerations of ethical implications. Addressing study limitations such as sample size constraints and data quality issues will be pivotal in enhancing the generalizability and applicability of predictive models across diverse healthcare settings. In summary, the integration of evolutionary feature selection methods marks a crucial advancement in geriatric healthcare, fostering more precise disease classification models that cater to the evolving needs of aging populations worldwide. These advancements hold promise for shaping future healthcare strategies, ultimately

enhancing the quality of life and health outcomes for elderly individuals globally.

Limitation

Collecting data from elderly individuals can be challenging due to age-related factors such as vision problems, speech difficulties, or cognitive impairments like dementia. These challenges necessitate using clear language, extra support during interviews or surveys, and adaptive methods to ensure elderly participants can comfortably engage and provide accurate information. Respecting their abilities and ensuring their comfort during the data collection process is crucial.

Acknowledgements

This research was supported by the Thailand Science Research and Innovation (TSRI) through a grant fund for Rajabhat Maha Sarakham University. Numerous advisors and researchers from the University of Phayao also provided partial support. The authors extend their gratitude to all for their support and cooperation in completing this research, facilitated by the Research and Development Institute (RDI) of Rajabhat Maha Sarakham University. Special thanks are extended for the research data support from the population in Kang Leung Chan Sub-district, Mueang District of Maha Sarakham Province.

References

[1] Ageing and Health, World Health Organization, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[2] Efraim Jaul, and Jeremy Barron, "Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population," *Frontiers in Public Health*, vol. 5, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] Pawel Posadzki et al., "Exercise/Physical Activity and Health Outcomes: An Overview of Cochrane Systematic Reviews," *BMC Public Health*, vol. 20, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Leonardo Santos Lopes da Silva et al., "Nutritional Status, Health Risk Behaviors, and Eating Habits are Correlated with Physical Activity and Exercise of Brazilian Older Hypertensive Adults: A Cross-Sectional Study," *BMC Public Health*, vol. 22, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jennifer E. Graham-Engeland et al., "Negative and Positive Affect as Predictors of Inflammation: Timing Matters," *Brain, Behavior, and Immunity*, vol. 74, pp. 222-230, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[6] Wongpanya Nuankaew, and Jaree Thongkam, "Improving Student Academic Performance Prediction Models Using Feature Selection," *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, Thailand, pp. 392-395, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rakibul Islam, Azrin Sultana, and Mohammad Rashedul Islam, "A Comprehensive Review for Chronic Disease Prediction Using Machine Learning Algorithms," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, pp. 1-28, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Girish Chandrashekar, and Ferat Sahin, "A Survey on Feature Selection Methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[9] Miguel García-Torres, Roberto Ruiz, and Federico Divina, "Evolutionary Feature Selection on High Dimensional Data Using a Search Space Reduction Approach," *Engineering Applications of Artificial Intelligence*, vol. 117, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Soo-Jin Lim et al., "Medical Health Records-Based Mild Cognitive Impairment (MCI) Prediction for Effective Dementia Care," *International Journal of Environmental Research and Public Health*, vol. 18, no. 17, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Erin Nitschke et al., "Impact of Nutrition and Physical Activity Interventions Provided by Nutrition and Exercise Practitioners for the Adult General Population: A Systematic Review and Meta-Analysis," *Nutrients*, vol. 14, no. 9, pp. 1-33, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Abbas Saad Alatrany et al., "An Explainable Machine Learning Approach for Alzheimer's Disease Classification," *Scientific Reports*, vol. 14, no. 1, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Munish Khanna et al., "An Enhanced and Efficient Approach for Feature Selection for Chronic Human Disease Prediction: A Breast Cancer Study," *Heliyon*, vol. 10, no. 5, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Junaid Rashid et al., "An Augmented Artificial Intelligence Approach for Chronic Diseases Prediction," *Frontiers in Public Health*, vol. 10, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Nina de Lacy, Michael J. Ramshaw, and J. Nathan Kutz, "Integrated Evolutionary Learning: An Artificial Intelligence Approach to Joint Learning of Features and Hyperparameters for Optimized, Explainable Machine Learning," *Frontiers in Artificial Intelligence*, vol. 5, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Preeti Khera, and Neelesh Kumar, "Ensemble Learning Classifier with Optimal Feature Selection for Parkinson's Disease," *2020 International Conference on Advances in Computing, Communication & Materials*, Dehradun, India, pp. 427-431, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yifan Qin et al., "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Abdullah Marish Ali, Farsana Salim, and Faisal Saeed, "Parkinson's Disease Detection Using Filter Feature Selection and a Genetic Algorithm with Ensemble Learning," *Diagnostics*, vol. 13, no. 17, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Prabhleen Kaur Chawla et al., "Parkinson's Disease Classification Using Nature Inspired Feature Selection and Recursive Feature Elimination," *Multimedia Tools and Applications*, vol. 83, pp. 35197-35220, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Suman Bhakar et al., "A Hybrid Model: Random Classification and Feature Selection Approach for Diagnosis of the Parkinson Syndrome," *Scalable Computing: Practice and Experience*, vol. 25, no. 1, pp. 167-176, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Sirage Zeynu, and Shruti Patil, "Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method," *WSEAS Transactions on Information Science and Applications*, vol. 15, pp. 168-176, 2018. [Google Scholar] [Publisher Link]

[22] Esty Purwaningsih, "Improving the Performance of Support Vector Machine with Forward Selection for Prediction of Chronic Kidney Disease," *Journal of Computer Science and Technology*, vol. 8, no. 1, pp. 18-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] K. Hema, K. Meena, and Ramaraj Pandian, "Analyze the Impact of Feature Selection Techniques in the Early Prediction of CKD," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 66-77, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jafar Abdollahi, and Babak Nouri-Moghaddam, "Feature Selection for Medical Diagnosis: Evaluation for Using a Hybrid STACKED-Genetic Approach in the Diagnosis of Heart Disease," *Arxiv*, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Peng Wang et al., "Multiobjective Differential Evolution for Feature Selection in Classification," *IEEE Transactions on Cybernetics*, vol. 53, no. 7, pp. 4579-4593, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Wongpanya S. Nuankaew, Sittichai Bussaman, and Pratya Nuankaew, "Evolutionary Feature Weighting Optimization and Majority Voting Ensemble Learning for Curriculum Recommendation in the Higher Education," *15th International Conference: Multi-Disciplinary Trends in Artificial Intelligence*, Virtual Event, pp. 14-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Afnan M. Alhassan, and Wan Mohd Nazmee Wan Zainon, "Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis," *IEEE Access*, vol. 9, pp. 87310-87317, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] F. Maulidina et al., "Feature Optimization Using Backward Elimination and Support Vector Machines (SVM) Algorithm for Diabetes Classification," *Journal of Physics: Conference Series*, *International Conference on Mathematics: Pure, Applied and Computation*, Surabaya, Indonesia (virtual), vol. 1821, pp. 1-8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Erin LeDell, and Sebastien Poirier, "H2O AutoML: Scalable Automatic Machine Learning," *7th ICML Workshop on Automated Machine Learning*, pp. 1-16, 2020. [Google Scholar] [Publisher Link]

[30] Harry Zhang, and Jiang Su, "Naive Bayes for Optimal Ranking," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 20, no. 2, pp. 79-93, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[31] Tianqi Chen, and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] T. Cover, and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967. [CrossRef] [Google Scholar] [Publisher Link]

[33] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986. [CrossRef] [Google Scholar] [Publisher Link]

[34] J.A. Nelder, and R.W.M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370-384, 1972. [CrossRef] [Google Scholar] [Publisher Link]

[35] Faris Hrvat, Lemana Spahić, and Amina Aleta, "Heart Disease Prediction Using Logistic Regression Machine Learning Model," *Proceedings of the Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON) and International Conference on Medical and Biological Engineering (CMBEBIH)*, Sarajevo, Bosnia and Herzegovina, vol. 1, pp. 654-662, 2024. [CrossRef] [Google Scholar] [Publisher Link]