scientific reports

OPEN

Hybrid metaheuristic optimization for detecting and diagnosing noncommunicable diseases

Saleem Malik[1]✉, S. Gopal Krishna Patro[2], Chandrakanta Mahanty[3], Saravanapriya Kumar[4]✉, Ayodele Lasisi[5], Quadri Noorulhasan Naveed[5], Anjanabhargavi Kulkarni[6], Abdulrajak Buradi[7], Addisu Frinjo Emma[8]✉ & Naoufel Kraiem[5]✉

In our data-driven world, the healthcare sector faces significant challenges in the early detection and management of Non-Communicable Diseases (NCDs). The COVID-19 pandemic has further emphasized the need for effective tools to predict and treat NCDs, especially in individuals at risk. This research addresses these pressing concerns by proposing a comprehensive framework that combines advanced data mining techniques, feature selection, and meta-heuristic optimization. The proposed framework introduces novel hybrid algorithms, including the Hierarchical Genetic Multiple Reduct Selection Algorithm (H-GMRA) and the Customized Function-based Particle Swarm Optimization with Rough Set Theory for NCD Feature Selection (CPSO-RST-NFS). These algorithms aim to address the challenges of feature selection, computational complexity, and disease classification accuracy. H-GMRA outperforms traditional methods by identifying minimal feature sets with high dependency ratios. CPSO-RST-NFS combines meta-heuristic optimization with feature selection, resulting in improved efficiency and accuracy. Through extensive experimentation on diverse NCD datasets, this research demonstrates the framework's ability to select informative features, improve classification accuracy, and contribute to better patient outcomes. By bridging the gap between computational efficiency and disease classification accuracy, this work offers valuable insights for healthcare practitioners and data analysts, ultimately advancing the field of NCD research. The proposed framework presents a significant step towards enhancing the early detection and management of NCDs, offering hope for more precise clinical predictions and improved patient care.

Keywords Non-Communicable diseases, Feature selection, Optimization algorithms

Data mining is a very important tool that has emerged in this data-rich age. Data mining has helped social networking, healthcare, and e-commerce acquire insights. These are crucial inputs for informed decision-making, notably in the healthcare sector, which has been stressed by COVID-19[1]. People with noncommunicable diseases, including heart disease and diabetes, the main causes of death worldwide, are at risk from fast- and slow-moving COVID-19 variations[2]. Healthcare practitioners are being bombarded with massive amounts of patient data, making early disease prediction difficult due to a lack of analytical tools. The project will demonstrate a method that allows healthcare providers to often evaluate patient data to speed up NCD diagnosis and treatment.

In earlier medical data investigations, classification algorithms were applied, but integration with high false alarm rates and low disease detection rates was difficult. Some research used preprocessing to reduce misclassifications, but most lacked accuracy and computing efficiency[3]. The main issue this study addresses is early NCD prediction utilizing patient data. We believe that combining computational techniques with feature selection optimization can greatly improve illness classification accuracy and speed[4]. To address the complexity of NCDs and their distinct research needs, the publication uses a new feature selection method. GAs and PSO

[1]Department of Computer Science and Engineering, P A College of Engineering, Mangalore, Karnataka, India. [2]School of Engineering, Sreenidhi University, Hyderabad, Telangana 501301, India. [3]Department of Computer Science & Engineering, GITAM School of Technology, GITAM Deemed to Be University, Visakhapatnam 530045, India. [4]Department of MCA, Sacred Heart College (Autonomous), Tirupattur 635601, Tamil Nadu, India. [5]Department of Computer Science, College Of Computer Science, King Khalid University, Abha, Saudi Arabia. [6]Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, India. [7]Nitte Meenakshi Institute of Technology, Bangalore, India. [8]College of Engineering and Technology, School of Mechanical and Automotive Engineering, Dilla University, Gedeo Zone, South Ethiopia Regional State, Po Box 419, Dilla, Ethiopia. ✉email: baronsaleem@gmail.com; priya@shctpt.edu; addisuf@du.edu.et; nkraiem@kku.edu.sa

are proven methods in this field, however the text acknowledges that one size does not fit all. This is why the core invention of this paper, the objective function, addresses the critical need for customisation. This modification is necessary to adjust algorithms to NCD data sets' complexity and demands. NCD research has unique problems, thus researchers must tailor feature selection algorithms to these challenges. Since customization is desired and important for handling NCDs and making feature selection more appropriate and accurate, the study emphasizes this[5]. Two new hybrid algorithms combine state-of-the-art and fresh notions.[6] For feature selection, a Hierarchical Genetic Multiple Reduct Selection Algorithm could outperform Quick Reduct Algorithms by ensuring minimal reducts with higher dependency ratios for better feature quality and more accurate disease classification[7]. We explore the integration of meta-heuristic algorithms with particle swarm optimization for feature selection. Customised Function-based Particle Swarm Optimization with Rough Set Theory for NCD Feature Selection—CPSO-RST-NFS. Feature selection and classification accuracy are improved over traditional QRA approaches[8].

We employ hybrid algorithms to balance exploration and exploitation for faster convergence and economical computing resource consumption in feature selection. Our research provides a solid foundation for NCD analysis and experimentation using several datasets. This clarifies NCD dataset properties and how they affect illness classification. We want to empower healthcare practitioners with powerful, data-driven NCD diagnosis and treatment tools to improve clinical forecasts and patient outcomes[9]. Healthcare practitioners and data analysts gain valuable insights and ideas on the bridge between computational efficiency and illness classification accuracy, which may lead to novel NCD research and management strategies. In this study, we use NCD datasets to demonstrate how our proposed approaches affect disease categorization outcomes. The results and contributions of this study could alter NCD diagnosis and management and give health practitioners stronger weapons against such lethal disorders[10].

Preliminaries

When evaluating their issue areas, the algorithms 'CPSO-RST-NFS' and 'H-GMRA' seem to be good contributions in feature selection and multiple reduct selection, respectively, but they are not novel. 'CPSO-RST-NFS' applies particle swarm optimization and rough set theory to feature selection, but it offers little new. Using genetic algorithms and other methods, 'H-GMRA' presents a Hierarchical Genetic Algorithm for Multiple Reduct Selection. Combining well-known methodologies adapted to challenges is innovative in this line of approaches, although the essential building blocks are well-established principles. Novelty lies in their adaptations and applications, not in new paradigms. Their practicality and efficiency in handling complicated environmental challenges make them important.

Heuristic methods

Problem-solving and decision-making are simplified with heuristic algorithms. Unlike slower algorithms that seek perfect solutions, these algorithms value speed over precision and completeness[11]. Optimizing problem-solving is their specialty. Heuristic algorithms help solve difficult, time-sensitive issues without precise solutions[12,13]. Greedy approaches prioritize locally optimal decisions, while local search methods constantly improve the answer.

Handling missing values

Missing data in well-structured studies may introduce bias and reduce the validity of the conclusions. Avoiding this requires data preprocessing. Pre-processing removes redundant, unnecessary, and noisy data to improve data quality. The current research uses mean imputation for missing values and min–max normalization for noise reduction in NCD datasets[14].After normalizing NCD datasets, feature selection follows. This crucial stage selects only relevant information, which the classifier then evaluates to predict diseases[15].

Rough set approach

This mathematical method, Rough Set Theory (RST)[16], can handle incomplete, imprecise, and inconsistent real-world datasets. It can extract structural correlations from noisy data, but best with discrete and binary data. RST requires discretization to ensure discernibility for continuous data. Similar information makes things indiscernible, generating elementary sets whose union may be crisp or rough, implying imprecision in RST[17]. According to, attribute reduction removes superfluous attributes from datasets to maintain only those relevant to high-quality classification, improving efficiency and interpretability. Conventional rough set procedures determine the optimal reduct, the minimal subset of attributes with classification accuracy equal to or greater than the original set using a classifier[18,19].

Quick reduct algorithm

The Quick Reduct Algorithm selects features in rough set theory[20]. It selects a selection of attributes to enhance classification accuracy and simplify high-dimensional datasets. By building a discernibility matrix, assessing attribute importance, and creating a minimal reduct, it retains important information and improves model efficiency and interpretability. Dimensionality reduction reduces dimensionality, improves efficiency, and reduces overfitting in large datasets[21].

Meta-heuristic optimization

Meta heuristic optimization ensures global optimality within a time restriction[22]. These methods balance intensification and diversification strategies in exploring solution spaces, with some being trajectory-based like Simulated Annealing and Hill Climbing and others population-based like PSO and Genetic Algorithms[23]. A new population-based PSO algorithm that uses a rough set-based filter mechanism and an objective function with the golden ratio to improve feature selection and classification of NCD datasets takes advantage of particles' iterative behavior to find optimal solutions[24].

Particle velocity update

The velocity of each particle ($v_i(t+1)$) at time t + 1 is updated using the following equation[25]:

$$V_i\,(t+1) = w*V_i\,(t) + C_1*rand_1*(pbest_i x - x_i\,(t)) + (2*rand_2*(gbest - x_i\,(t)) \quad (1)$$

Above, w is the inertia weight; c_1 and c_2 are the cognitive and social learning factors; $rand_1()$ and $rand_2()$ generate random values in the range [0, 1]; $pbest_i$ refers to the personal best position of the i^{th} particle; and gbest is the global best position.

Particle position update

The position of each particle $(x_i(t+1))$ at time t + 1 is updated based on its velocity[26]:

$$\mathrm{x_i\,(t+1) = x_i\,(t) + v_i\,(t+1)} \quad (2)$$

These equations govern how particles explore the solution space and interact with each other to find optimal solutions. The selection of these equations and parameters is driven by the need to adapt the optimization process to the characteristics of our research domain.

Amalgamation of particle swarm optimization and rough set theory

This study develops a PSO algorithm with a customized mathematical function utilizing Rough Set Theory to handle data dimensionality and feature selection difficulties. The approach guides PSO to search for a restricted and meaningful group of attributes to maximize feature selection and classification in NCD datasets. Offshoots of feature selection improve efficiency and efficacy, with almost original prognostic capability and reduced computing complexity of heuristic and exhaustive search strategies[27]. RST is crucial to feature selection in the CPSO-RST-NFS algorithm because it detects significant properties from accessed data dependencies. While RST encompasses a good number of equations and steps, one important component of this technique is the computation of the dependency ratio for the selected features, given by the equation[28]:

$$Gamma = \frac{(|X' - X|)}{|X|} \quad (3)$$

Here, X' represents the reduct, which is the minimal subset of features, and X represents the complete set of features. The dependency ratio Gamma quantifies the relevance of the selected features for classification tasks. The integration of PSO and RST within CPSO-RST-NFS combines the rapid convergence and simplicity of PSO with the data dependency analysis capabilities of RST. This synergy empowers the algorithm to efficiently select feature subsets that enhance classification accuracy in NCD datasets[29].

Literature review

The Literature Review is crucial to research since it provides insight into past studies in that topic. It can give aspiring researchers an overview of recent advances in their field. This section will discuss the research on hybrid algorithms and meta-heuristics used in NCDs to improve patient illness prediction. Rough set-based feature selection algorithms, based on rough set theory[30], efficiently identify and select critical characteristics from high-dimensional data sets[31].Dependency measures, relevance, reducts, heuristic methods to reduce classification errors, ranking features by relevance, and hybridization of RS techniques with other optimization algorithms for improved selection are all used[32]. Since discernibility and dependency minimize dimensionality while keeping significant information, these methods are effective for machine learning feature selection, especially with large and complicated datasets. Table 1: Related hybrid techniques summary[33] Introduced the variable precision neighborhood rough set decision system, which uses neighborhood granular swarm[34], variable precision approximation sets[35] and positive regions to pick feature subsets based on attribute significance. While using SVM on UCI datasets, our approach showed good feature recognition and defect tolerance. Rough set approaches work well for smaller datasets. A Spark framework[36] using distributive rough set techniques for feature selection on huge datasets[37] addressed this constraint. This method divides the dataset into smaller groups for parallel processing, reducing features but increasing processing time, especially for bigger datasets.

Several researchers have developed novel rough set-based feature selection methods.[38] Developed a rough set-based Ant Lion optimizer to find minimal reducts[39] and refined a rough set theory-based breast cancer feature selection method[40].

Heuristic algorithms seek global optima by balancing exploitation and exploration[48]. PSO has been used by many studies (Table 2) to find global optima, typically with other metaheuristic methods. Various research projects have focused on hybrid PSO algorithms. A filter-based strategy using the Whale Optimization Algorithm (WOA) with a Kaggle dataset was used to diagnose breast cancer in[49]. Extremely Randomized Tree (ERT), SVM, KNN, NB, Stochastic Gradient Descent (SGD), RF, DT, LR, and Kernel SVM classifiers were used in Python experiments. Evaluation metrics were accuracy, recall, F-Score, and precision. The program outperformed other classifiers with a 0.7% accuracy gain when employing ERT. In another context,[50] introduced a Hybrid PSO with the Cuckoo Search Algorithm for complex nonlinear engineering issues. This algorithm efficiently approximated global optima, proving its problem-solving abilities.[51] also introduced Enhanced Partial Search Particle Swarm Optimization, a cooperative multi-swarm strategy to improve PSO search behavior. By actively exploring global

Authors	Data Source	Algorithm Used	Advantages	Limitations
[41]	UCI	Variable Precision Neighborhood Rough Set Decision System	- Effective feature identification.- Fault tolerance.- Utilized SVM for evaluation	- Rough set methods may be less suitable for large datasets.- Limited to UCI datasets
[42]	UCI	Spark framework with distributive rough set approaches	- Addresses scalability with large datasets through parallel processing.- Reduces features effectively	- Increased processing time for larger datasets
[43]	UCI	Ant Lion optimizer	- Proposed a novel rough set-based feature selection method.- Identifies minimal reducts	- Lack of dataset and context details.- Limited evaluation and comparison with other methods
[44]	Breast cancer data	Rough set theory-based feature selection algorithm	- Tailored for breast cancer classification.- Utilizes rough set theory for feature selection	- Specific to breast cancer classification.- May not apply to other types of datasets
[45]	UCI	Heuristic-based feature selection within the rough set framework	- Utilizes heuristics and decision rules for feature subset selection.- Focus on maintaining rule quality	- Lack of dataset and context details.- Limited evaluation on specific applications
[46]	UCI	Greedy algorithm for multiple reducts	- Focuses on unique attribute selection to enhance dataset optimality.- Explores multiple reducts	- Lack of dataset and context details.- May not generalize to various domains and datasets
[47]	UCI	Modified quick reduct algorithm	- Reduces the size of information systems horizontally.- Eliminates objects in lower approximation	- Lack of dataset and context details.- Limited evaluation on specific applications.- May not be suitable for all datasets

Table 1. Comparative Overview of Studies on Hybrid algorithms in NCD datasets.

Authors	Data Source	Algorithm Used	Advantages	Limitations
[48]	Cagle breast cancer dataset	Whale Optimization Algorithm (WOA)	- Improved breast cancer detection accuracy compared to other classifiers	- Limited information about the dataset
[49]	UCI	Hybrid PSO and Cuckoo Search Algorithm	- Effectively approximates global optima	- Lacks information about the dataset
[50]	UCI	Enhanced Partial Search Particle Swarm Optimization	- Mitigates the risk of particles getting trapped in local optima.- Directs exploration towards global optima	- No specific dataset or context mentioned
[52]	UCI	Hybrid PSO and Harmony Search	- Consistent feature selection outcomes	- Longer processing durations
[53]	UCI	Binary Genetic Swarm Optimization with PSO	- Integrates local exploration to refine solutions.- Employs KNN and MLP as classifiers	- Limited information about the dataset
[54]	UCI	Wrapper-centric framework with bio inspired algorithms	- Diminishes feature subsets with high precision	- No specific dataset or context mentioned
[55]	UCI	Hybrid feature selection with PSO and Symmetrical Uncertainty (SU)	- Amplifies classification accuracy.- Reduces computational time	- No specific dataset or context mentioned
[56]	Gene expression datasets	Bare-Bone Particle Swarm Optimization (BBPSO)	- Enhances feature discretization in gene expression datasets.- Augments KNN classifier performance	- Specific to gene expression datasets
[57]	Indian pines and Toronto datasets	Hybrid PSO and GA with SVM	- Improves road identification accuracy in remote sensing data.- Works within CPU processing constraints	- Limited to remote sensing datasets
[58]	Various datasets including Breast Cancer, Diabetes, and Hepatitis	Weighted least square technique with SVM and feature selection	- Mitigates class imbalance issues.- Enhances classifier accuracy	- Dataset-specific approaches
[39]	UCI	PSO-based single and multi-objective strategies	- Utilizes multiple classifiers and optimization strategies	- No specific dataset or context mentioned
[59]	UCI	Wrapper-driven approach fusing PSO and correlation-centered feature selection	- Optimizes classification precision across various algorithms	- No specific dataset or context mentioned
[60]	UCI	Asynchronous PSO	- Ensures convergence through dynamic particle behavior	- No specific dataset or context mentioned
[61]	UCI	Binary PSO with Mutual Information	- Utilizes entropy and mutual information for feature selection	- No specific dataset or context mentioned
[62]	UCI	Hybrid PSO and Differential Evolution	- Designed for optimizing functions	- No specific dataset or context mentioned
[38]	UCI	Hybrid PSO and GA (PSOGA)	- Swift convergence for multimodal function optimization	- No specific dataset or context mentioned
[63]	UCI	Hybrid PSO and GSA	- Combines social thinking and search capabilities	- Challenges with local optima and algorithmic complexity due to uniform random variables
[64]	UCI	Hybrid GSASVM	- Efficient feature selection and improved classification precision	- No specific dataset or context mentioned

Table 2. Comparative Overview of Studies on Heuristic Algorithms in NCD datasets.

optima, this method reduces the danger of particles being stuck in local optima. This method is useful for unconstrained optimization."

[52] Combined PSO and Harmony Search for feature selection, producing consistent results but longer processing times.[53] Proposed a binary genetic swarm optimization-PSO combination using KNN and MLP classifiers and local exploration to refine solutions.[54] Presented a wrapper-centric framework using correlation-based ensemble feature selection, bio-inspired algorithms including Differential Evolution, Lion Optimization, and Glow Worm Swarm Optimization, and AdaBoostSVM for classification. This strategy reduced high-

precision feature subsets. A hybrid feature selection strategy combining PSO and Symmetrical Uncertainty (SU) can improve classification accuracy and save processing time when combined with Naïve Bayes and J48[55],[56] Used KNN as a classifier in Bare-Bone Particle Swarm Optimization (BBPSO) to discretize gene expression dataset characteristics.[57] Introduced a hybrid PSO and GA technique using SVM for road identification in Indian pines and Toronto datasets to improve classification accuracy within CPU processing restrictions[58] Fine-tuned features in Breast Cancer, Diabetes, and Hepatitis datasets using weighted least square with SVM, correlation-based feature selection, and Sequential Forward Selection (SFS). This method reduced class imbalance and improved classifier accuracy[39]. Introduced PSO-based single and multi-objective techniques, combining SVM, KNN, Decision Trees, and Naïve Bayes with mutual information and entropy[59]. Developed a wrapper-driven technique combining PSO and correlation-centered feature selection to enhance classification accuracy in various algorithms, such as Naïve Bayes, Decision Trees, and C4.5, RBF, KNN, and Bayesian classifiers[60]. Ensured convergence in asynchronous PSO via dynamic particle behavior. Entropy-rooted Binary PSOs with Mutual Information were used with decision trees to choose features[61,62]. Optimised functions using a hybrid PSO/DE algorithm[38]. proposed a fast-convergent hybrid PSO-GA (PSOGA) for multimodal function optimization[63]. used PSO and GSA in a hybrid method, combining social reasoning with search, but faced local optima and uniform random variable algorithmic complication. Provided a hybrid GSASVM algorithm for efficient feature selection and improved classification precision.

The requirement for empirical validation in feature selection is well-established, but this research emphasizes customization, especially for non-communicable diseases. This work's main novelty is adapting and customizing feature selection algorithms for NCD research. Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO) are feature selection methods, but we focus on customizing them for NCD datasets. These chronic disease patient datasets differ from machine learning datasets due to their many variables, complex nonlinear interactions, and nuanced patterns[41]. These dataset-specific complications suggest that present feature selection approaches may not be suitable for NCDs. Empirical proof is essential to research, but a complete comparison analysis spanning NCD-related datasets and disorders is logistically burdensome. We believe this research's strategy, which stresses NCD-specific feature selection strategies, is feasible and effective. It recognizes that NCD study requires appropriate feature selection to maximize accuracy in their unique setting.

Proposed frameworks

This paper proposes a two-phase architecture to improve NCD classification accuracy and efficiency in Fig. 1. H-GMRA, the initial phase in this work, selects features from an input dataset with substantial NCD attribute variation. Thus, feature selection precedes data preparation to ensure data quality[65]. QRA systematically analyzes attribute space and generates many reducts that can be used as bases. Additionally, each reduct's dependency values are determined to quantify their relevance to NCD classification. H-GMRA uses the QRA-generated reduct to identify minimum features with high dependency values, using the threshold dependency value

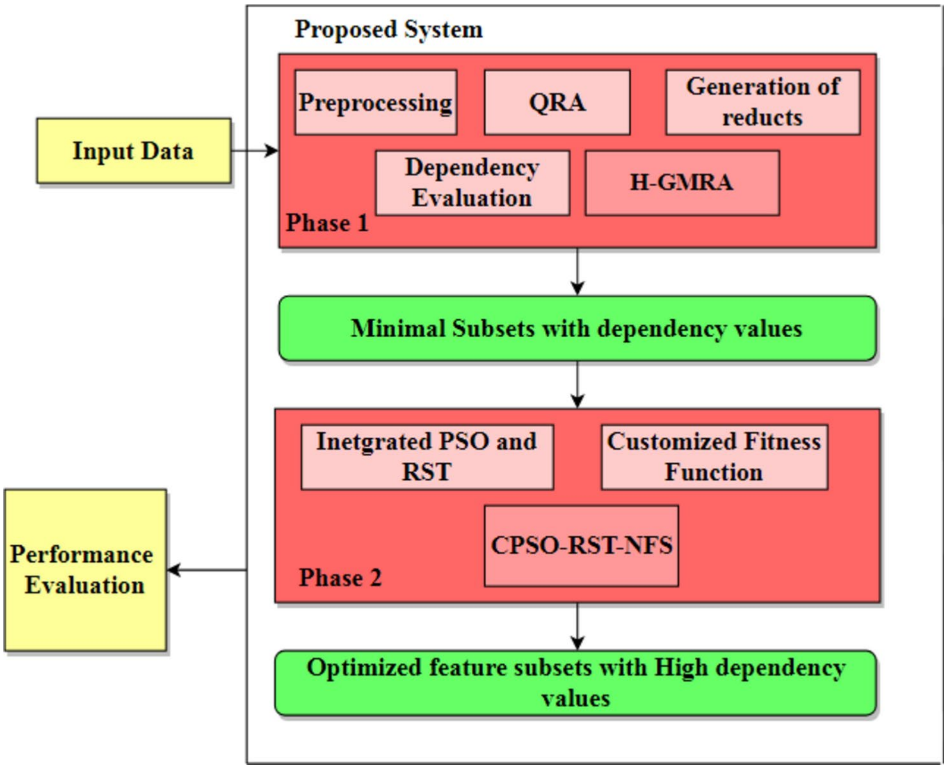
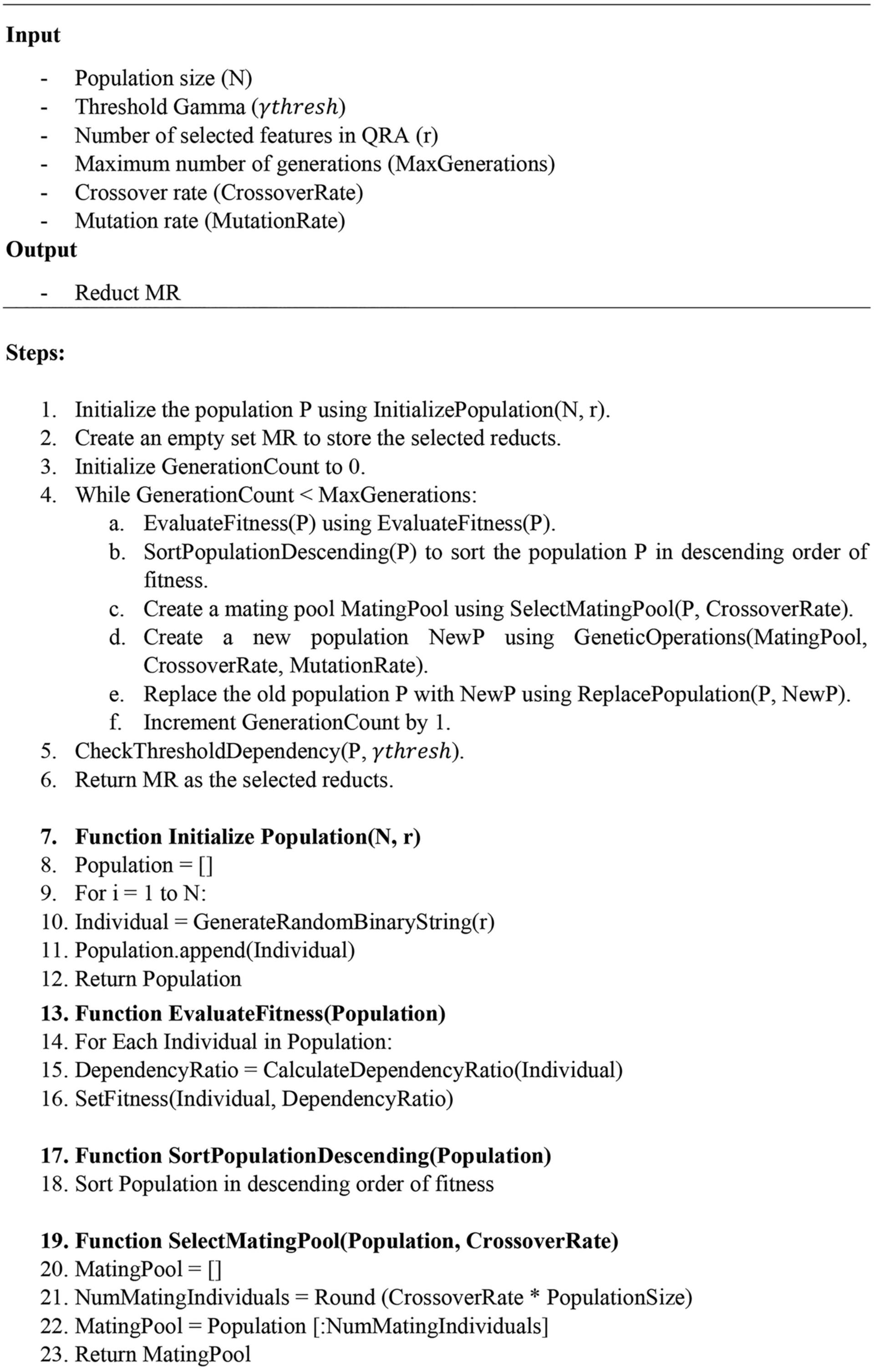


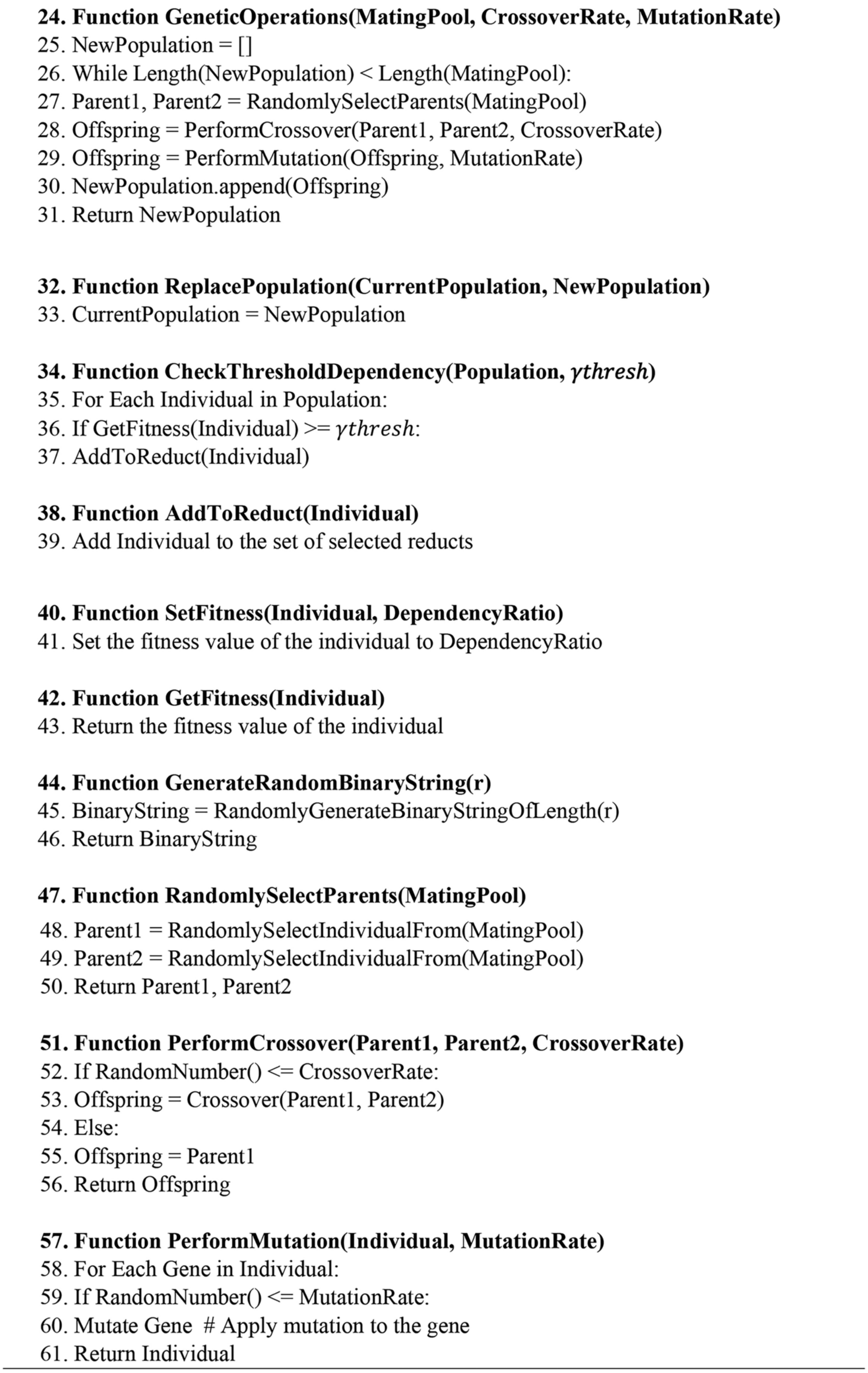
Fig. 1. Detailed View of proposed system.



```
Input
    -   Population size (N)
    -   Threshold Gamma ($\gamma thresh$)
    -   Number of selected features in QRA (r)
    -   Maximum number of generations (MaxGenerations)
    -   Crossover rate (CrossoverRate)
    -   Mutation rate (MutationRate)
Output
    -   Reduct MR

Steps:

    1.  Initialize the population P using InitializePopulation(N, r).
    2.  Create an empty set MR to store the selected reducts.
    3.  Initialize GenerationCount to 0.
    4.  While GenerationCount < MaxGenerations:
            a.  EvaluateFitness(P) using EvaluateFitness(P).
            b.  SortPopulationDescending(P) to sort the population P in descending order of
                fitness.
            c.  Create a mating pool MatingPool using SelectMatingPool(P, CrossoverRate).
            d.  Create a new population NewP using GeneticOperations(MatingPool,
                CrossoverRate, MutationRate).
            e.  Replace the old population P with NewP using ReplacePopulation(P, NewP).
            f.  Increment GenerationCount by 1.
    5.  CheckThresholdDependency(P, $\gamma thresh$).
    6.  Return MR as the selected reducts.

    7.  Function Initialize Population(N, r)
    8.  Population = []
    9.  For i = 1 to N:
    10. Individual = GenerateRandomBinaryString(r)
    11. Population.append(Individual)
    12. Return Population
    13. Function EvaluateFitness(Population)
    14. For Each Individual in Population:
    15. DependencyRatio = CalculateDependencyRatio(Individual)
    16. SetFitness(Individual, DependencyRatio)

    17. Function SortPopulationDescending(Population)
    18. Sort Population in descending order of fitness

    19. Function SelectMatingPool(Population, CrossoverRate)
    20. MatingPool = []
    21. NumMatingIndividuals = Round (CrossoverRate * PopulationSize)
    22. MatingPool = Population [:NumMatingIndividuals]
    23. Return MatingPool
```

Algorithm 1. Hierarchical Genetic Multiple Reduct Selection Algorithm (H-GMRA).

($\gamma thresh$) as the selection criterion[66]. Phase 2 will get subsets of reduced features with high dependency values from Phase 1. The Phase 2 algorithm is CPSO-RST-NFS. This step starts with step 1's reduced feature subsets. A population of particles, each representing a subset of features, is initialized by the PSO component. It evaluates particle fitness using a customized fitness algorithm based on subset feature counts and dependence values. PSO iteratively refines subsets of features, altering particle positions based on fitness and how far from the best each particle is[67]. Thus, Phase 2 will generate upgraded feature subsets with higher classification accuracy and computing efficiency. After CPSO-RST-NFS optimization, we compare the subsets' performance to QRA

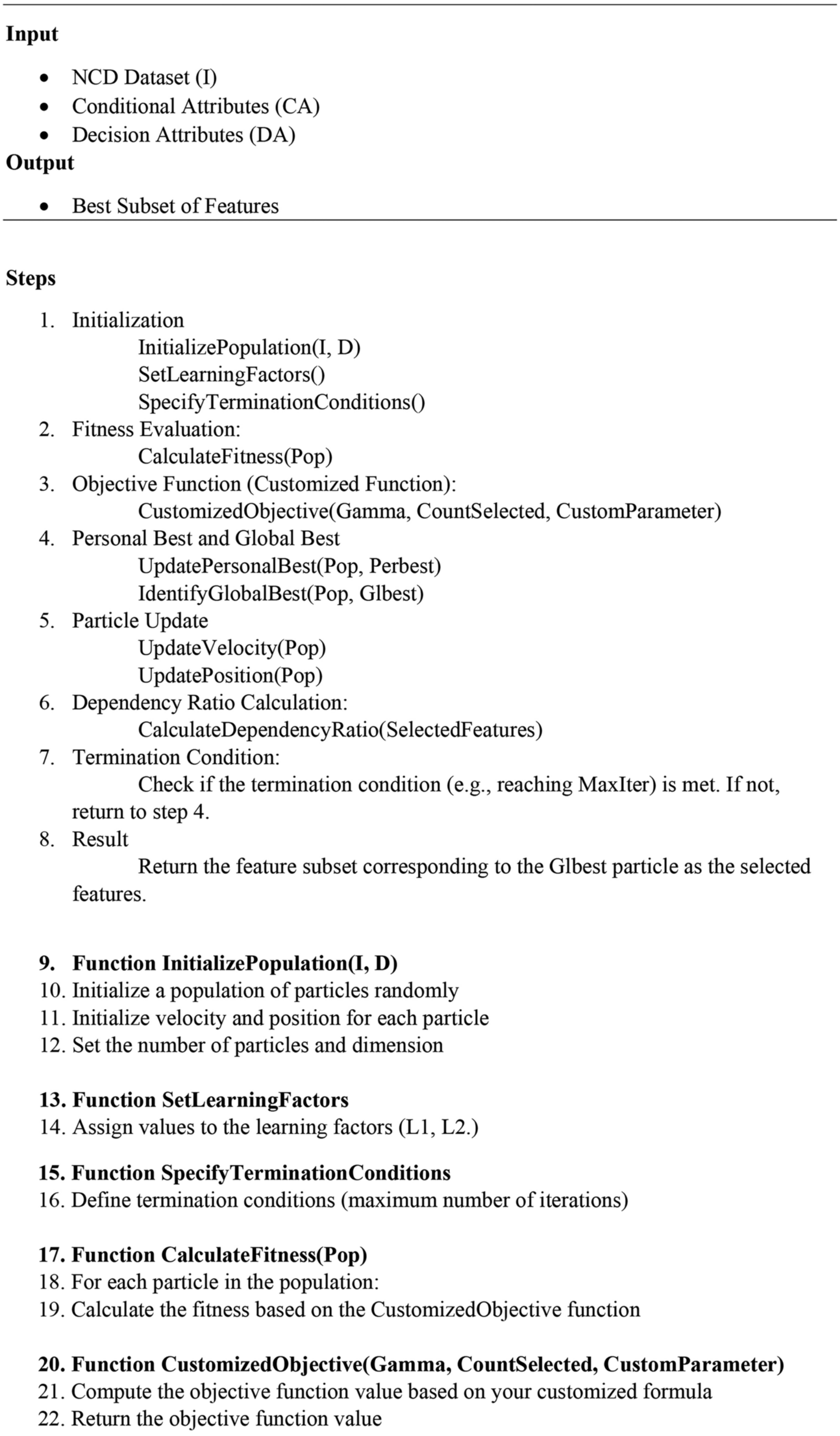


```
Function GeneticOperations(MatingPool, CrossoverRate, MutationRate)
    NewPopulation = []
    While Length(NewPopulation) < Length(MatingPool):
        Parent1, Parent2 = RandomlySelectParents(MatingPool)
        Offspring = PerformCrossover(Parent1, Parent2, CrossoverRate)
        Offspring = PerformMutation(Offspring, MutationRate)
        NewPopulation.append(Offspring)
    Return NewPopulation

Function ReplacePopulation(CurrentPopulation, NewPopulation)
    CurrentPopulation = NewPopulation

Function CheckThresholdDependency(Population, $\gamma thresh$)
    For Each Individual in Population:
        If GetFitness(Individual) >= $\gamma thresh$:
            AddToReduct(Individual)

Function AddToReduct(Individual)
    Add Individual to the set of selected reducts

Function SetFitness(Individual, DependencyRatio)
    Set the fitness value of the individual to DependencyRatio

Function GetFitness(Individual)
    Return the fitness value of the individual

Function GenerateRandomBinaryString(r)
    BinaryString = RandomlyGenerateBinaryStringOfLength(r)
    Return BinaryString

Function RandomlySelectParents(MatingPool)
    Parent1 = RandomlySelectIndividualFrom(MatingPool)
    Parent2 = RandomlySelectIndividualFrom(MatingPool)
    Return Parent1, Parent2

Function PerformCrossover(Parent1, Parent2, CrossoverRate)
    If RandomNumber() <= CrossoverRate:
        Offspring = Crossover(Parent1, Parent2)
    Else:
        Offspring = Parent1
    Return Offspring

Function PerformMutation(Individual, MutationRate)
    For Each Gene in Individual:
        If RandomNumber() <= MutationRate:
            Mutate Gene  # Apply mutation to the gene
    Return Individual
```

Algorithm 1. (continued)

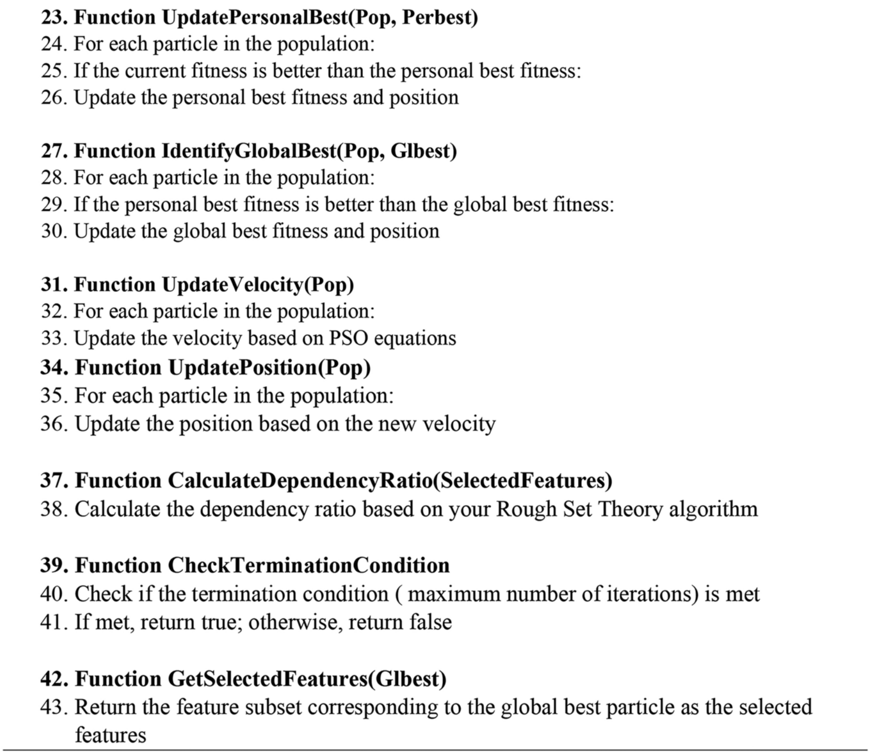
subsets. The research will assist address all NCD classification issues and improve clinical forecasts and patient recovery.

Two new feature selection algorithms are described in Sections "Selection" and "Genetic Operations". First, we will introduce the Hierarchical Genetic Multiple Reduct Selection Algorithm (H-GMRA), which uses genetic algorithms and rough set theory to pick optimal features. H-GMRA will use a new parameter, threshold dependency ratio, to ensure feature relevance. The difficulty of high-dimensional datasets is widespread in current, data-driven research. CPSO-RST-NFS, a unique NCD feature selection technique, combines PSO



```
Input
    • NCD Dataset (I)
    • Conditional Attributes (CA)
    • Decision Attributes (DA)
Output
    • Best Subset of Features

Steps
 1. Initialization
            InitializePopulation(I, D)
            SetLearningFactors()
            SpecifyTerminationConditions()
 2. Fitness Evaluation:
            CalculateFitness(Pop)
 3. Objective Function (Customized Function):
            CustomizedObjective(Gamma, CountSelected, CustomParameter)
 4. Personal Best and Global Best
            UpdatePersonalBest(Pop, Perbest)
            IdentifyGlobalBest(Pop, Glbest)
 5. Particle Update
            UpdateVelocity(Pop)
            UpdatePosition(Pop)
 6. Dependency Ratio Calculation:
            CalculateDependencyRatio(SelectedFeatures)
 7. Termination Condition:
            Check if the termination condition (e.g., reaching MaxIter) is met. If not,
    return to step 4.
 8. Result
            Return the feature subset corresponding to the Glbest particle as the selected
    features.

 9. Function InitializePopulation(I, D)
10. Initialize a population of particles randomly
11. Initialize velocity and position for each particle
12. Set the number of particles and dimension

13. Function SetLearningFactors
14. Assign values to the learning factors (L1, L2.)

15. Function SpecifyTerminationConditions
16. Define termination conditions (maximum number of iterations)

17. Function CalculateFitness(Pop)
18. For each particle in the population:
19. Calculate the fitness based on the CustomizedObjective function

20. Function CustomizedObjective(Gamma, CountSelected, CustomParameter)
21. Compute the objective function value based on your customized formula
22. Return the objective function value
```

Algorithm 2. Customized Function-based Particle Swarm Optimization with Rough Set Theory for NCD Feature Selection (CPSO-RST-NFS).



```
Function UpdatePersonalBest(Pop, Perbest)
    For each particle in the population:
        If the current fitness is better than the personal best fitness:
            Update the personal best fitness and position

Function IdentifyGlobalBest(Pop, Glbest)
    For each particle in the population:
        If the personal best fitness is better than the global best fitness:
            Update the global best fitness and position

Function UpdateVelocity(Pop)
    For each particle in the population:
        Update the velocity based on PSO equations
Function UpdatePosition(Pop)
    For each particle in the population:
        Update the position based on the new velocity

Function CalculateDependencyRatio(SelectedFeatures)
    Calculate the dependency ratio based on your Rough Set Theory algorithm

Function CheckTerminationCondition
    Check if the termination condition ( maximum number of iterations) is met
    If met, return true; otherwise, return false

Function GetSelectedFeatures(Glbest)
    Return the feature subset corresponding to the global best particle as the selected
    features
```

Algorithm 2. (continued)

and RST. It balances classification performance and computing economy using a multi-objective function[69]. These include a customizable goal function that adapts to research and problem domains. Its focus on non-communicable illness research makes CPSO-RST-NFS a unique healthcare contribution.

Phase 1: H-GMRA

The current data-rich environment requires feature selection to avoid overfitting in machine learning models by selecting critical characteristics, enhancing model accuracy and minimizing dimensionality. Supervised, unsupervised, and semi-supervised feature selection methods are used in this work. Filter, wrapper, and embedding models use Forward Selection, Backward Elimination, and Heuristic-Based Selection to remove redundant features during supervised feature selection[50].

H-GMRA algorithm and its components

The H-GMRA optimizes feature selection in high-dimensional datasets using rough set theory genetic algorithms. H-GMRA iteratively selects relevant features and excludes irrelevant attributes using a threshold dependence ratio to improve classification accuracy. This method navigates feature subset complexity, making it ideal for high-dimensional data applications. So, these basic component parts of this algorithm may be further broken down into:

Input parameters

- Population Size (N): The number of individuals in the population, each representing a potential subset of features.
- Threshold Gamma (γ*thres*h): A predefined threshold for the dependency ratio, ensuring that selected reducts meet a certain level of dependency.
- Number of Selected Features in Quick Reduct Algorithm (r): The number of features selected by the Quick Reduct Algorithm.
- Maximum Number of Generations (MaxGenerations): The maximum number of iterations the algorithm will run.
- Crossover Rate (CrossoverRate): The probability of applying crossover during genetic operations.
- Mutation Rate (MutationRate): The probability of applying mutation during genetic operations.

Initialization

Initialize a population P with N individuals. Each individual is represented as a string of bits of length r so that 1 stands for the presence of a feature, while 0 corresponds to its absence. This means the initialization of the population would be random.

Fitness assessment

Evaluate the fitness of every individual in the population according to (γ) for DA. At this step, the extent by which every individual will contribute towards the power of discernment of the objects in the dataset will be calculated.

Selection

Sort the population in decreasing order of fitness, making sure that those having higher dependency ratios are in the front. After that, choose the best of the individuals to enter into a mating pool where the percentage of the best selected is based on the Crossover Rate.

Genetic operations

Create a new population (NewP) by applying genetic operations (crossover and mutation) to the individuals in the mating pool. Crossover combines attributes from two parents, while mutation introduces small changes in an individual's attributes. The probability of applying crossover and mutation is determined by the CrossoverRate and MutationRate, respectively.

Replacement

Replace the old population (P) with the new population (NewP), continuing the evolutionary process.

Iterative refinement

Repeat the evaluation, selection, genetic operations, and replacement steps for a predefined number of generations (MaxGenerations). This iterative process allows the algorithm to explore different combinations of features.

Threshold dependency check

After the specified number of generations, filter the final population based on the threshold dependency ratio (γ*thres*h). Individuals whose dependency ratios meet or exceed this threshold are considered selected reducts.

Output

Return the selected reducts (MR) as the final result of the algorithm, representing the most relevant subset of features.

The Hierarchical Genetic Multiple Reduct Selection Algorithm (H-GMRA) selects the most important dataset properties. The approach uses evolutionary algorithms to iteratively select features that enhance classification accuracy while considering a threshold dependence ratio. Binary encoded individuals are used to start this process, and their dependency ratios determine their fitness. The greatest Matting pool is chosen to undergo crossover and mutation to generate a new population. It updates generations till the limit is reached. Finally, selected reductions will include people with dependency ratios above the threshold. H-GMRA reduces dimensionality systematically making it suited for complicated data sets and classification tasks.

Phase 2: CPSO-RST-NFS

The rise of new infectious diseases challenges medical prediction and diagnosis. Most NCDs are linked to other health issues. Removing unnecessary and superfluous attributes improves prediction accuracy. Basic feature selection for disease diagnosis is done using heuristic methods. These algorithms specialize in certain fields or problems because they solve different datasets differently. We will provide meta-heuristic techniques that are adaptable and random to address this issue. Filter-based feature selection solves challenges distinct from the induction process in a straightforward, rapid, and impartial manner[50]. Evolution-inspired algorithms were created by researchers[47].

CPSO-RST-NFS algorithm and its components

This paper proposes Algorithm 2 (CPSO-RST-NFS) to efficiently choose NCD dataset features. The method seeks to choose only important data features without losing integrity and improve classification accuracy. It does this by combining PSO and Rough Set Theory to their advantage. Rough Set Theory discovers data dependencies to define meaningful feature subsets, while PSO provides rapid convergence and simplicity for feature selection. It works with a Multi-Objective Function to balance fitness and feature count like a bi-objective function to optimize classification performance and computing efficiency[69]. Due to its considerable customization potential, CPSO-RST-NFS was chosen to ensure data quality and optimize NCD dataset classification outcomes.

The CPSO-RST-NFS algorithm consists of several components that work together to perform feature selection in the context of NCD datasets. Let's break down the key components of this algorithm:

Initialization

- InitializePopulation(I, D): This component initializes the population of particles. In the context of feature selection, each particle represents a potential subset of features.
- SetLearningFactors(): Learning factors, such as L1 and L2, are assigned to control how particles explore the solution space during optimization.
- SpecifyTerminationConditions(): Termination conditions are defined to determine when the algorithm should stop. Common conditions include reaching a maximum number of iterations or achieving a specific fitness threshold.

Fitness evaluation (CalculateFitness)

- This component calculates the fitness value for each particle in the population. Fitness represents how well a particular subset of features performs in terms of classification accuracy. The objective function, which incorporates the customized function, is used for fitness evaluation.

Objective function (Customized Function)

- The objective function is a crucial part of the algorithm. It quantifies the quality of a feature subset. It considers various factors, including the dependency ratio (Gamma), the count of selected attributes, and a custom parameter. The objective function guides the optimization process by assigning fitness values to particles.

Personal best and global best

- UpdatePersonalBest(Pop, Perbest): This component updates the personal best positions for each particle. Personal best positions represent the best solution encountered by each particle so far.
- IdentifyGlobalBest(Pop, Glbest): The global best particle is identified from the population. It represents the overall best solution found by any particle in the entire population.

Particle update

- UpdateVelocity(Pop): This step updates the velocity of each particle based on its current position, personal best position, and global best position. Velocity controls how particles explore the solution space.
- UpdatePosition(Pop): The particles' positions are updated based on their velocities. Based on this update, a new subset of features is determined for every particle.

Dependency ratio calculation: calculatedependencyratio(SelectedFeatures)
This calculation uses the concept of Rough Set Theory to compute the dependency ratio that gives Gamma for the selected features. This ratio comes in handy when assessing if the chosen features are relevant for classification.

Termination condition

- The algorithm checks termination conditions to decide whether to continue or stop the optimization process. Common conditions include reaching a maximum number of iterations or achieving a predefined fitness threshold.

Result

After termination, the method returns the global best particle feature subset. Select features improved for classification accuracy are in this subgroup. Together, these components help the program choose a suitable NCD dataset subset. The algorithm optimizes the goal function and considers numerous aspects to improve classification accuracy while minimizing features, making it useful for feature selection in healthcare and NCD research.

Objective function customization

Our feature selection methodology's objective function customization is innovative and important. Traditional feature selection algorithms include GAs and PSO[70]. Our technique is unique because we deliberately and problem-tailor the objective function. Customization takes NCD requirements and restrictions into account. A basic goal function balances the quantity of selected features with the classification model's accuracy. In NCD research, feature selection optimization can greatly affect algorithm efficiency[71]. In our optimization problem, the objective function can be built according to domain requirements and restrictions, but it may also need to be customized.

We now introduce a custom-made objective function to serve as a means through which we can inject mathematical insight into the optimization process.

The general form of our objective function we define as:

$$objectivefunction = a * \gamma - b * \frac{R}{D} + f\left(x\right) \tag{4}$$

- *a* and *b* are constants.
- *γ* (Gamma) is the dependency ratio calculated using Rough Set Theory.
- *R* is the sum of resulting attributes.
- *D* is the total of conditional attributes.
- *f(x)* can be any user-defined mathematical function depending on the need of your research.

It will be necessary to introduce a custom mathematical function, denoted as f(x), so that certain characteristics of our optimization problem can be accounted for. This will enable the tuning of the objective function into conformance with our research objectives.

$$f(x) = \sin(x) \tag{5}$$

x can be an angle or any other optimization issue parameter. The sine function makes the goal function periodic, which may help in optimization difficulties. Our customized objective function can be tailored to our research domain wants and desires. This will let us study how periodicity affects optimization, which may be interesting in some circumstances. Applying the angle parameter $x = \pi / 4$ (45 degrees in radians) is a reliable choice for several optimization objectives. First, it's a common angle in mathematics and engineering, so readers can understand it. A second argument is that radians simplify the trigonometric element of our target function and follow mathematical convention. This clarifies and standardizes a universal service and can be tweaked to improve our approach. Additionally, x = π/4 may be relevant to our research domain, explaining its selection. Our optimization strategy is based on the strong foundation of mathematical clarity, adaptability for testing, and possible domain relevance of x = π/4.

Objective function customization has two roles in our research. It will adjust the optimization technique to NCD datasets' unique demands. This will also show how periodicity affects optimization in a dimension that may be useful. Our methodology, specialized to NCD research, relies on objective function customization to strike this delicate balance between feature count and excellent accuracy.

Experimental setup

NCDs are one class of non-communicable chronic diseases mainly represented by Cardiovascular Disease, Diabetes, Cancer, and Liver Disease, responsible for 71% of the world's deaths according to the World Health Organization[36]. In the present study, the above four are focused on regarding NCD categories. The proposed algorithm is implemented in MATLAB 2016a on a system with an i5 processor, 64-bit Windows 8 OS, operating at 2.60 GHz, and equipped with 4 GB of RAM. For this study, it has selected those datasets corresponding to the NCD categories from the UC Irvine Machine Learning Repository, UCI. In this regard, each dataset was fed individually to the proposed feature selection algorithms and their accuracies were compared in disease diagnosis. Here, it is important to mention that all datasets used in this study are binary datasets whose target variables have two different classes: 0 for a healthy person and 1 for a patient. These datasets have represented a broad spectrum of NCDs and data types, thus allowing for the evaluation of the efficacy of H-GMRA in feature selection and disease diagnosis for various healthcare applications. Table 3 gives the Summary of datasets.

Classification techniques

These three different classifiers are implemented in this work: Support Vector Machine (SVM)[80], Decision Trees (DT)[33], and Naïve Bayes (NB)[23]. Each one of them is very different and appropriate to different types of datasets or feature selection methods.

Support vector machine

Due to its efficiency in high-dimensional data and complex decision boundaries, SVM is suitable for datasets with multiple features. SVM also works with linear and nonlinear data, making it flexible for varied NCD datasets. It handles skewed datasets well, which is common in health data where some diseases are rare.

Decision trees

Decision Trees are straightforward to read and can explain which features classify diseases best. They can handle categorical and continuous data, which is essential for clinical datasets with multiple attribute types. They can also find non-linear correlations between characteristics and outcomes, enhancing SVM performance.

Naïve bayes

The research utilizes Naïve Bayes, a computationally efficient probabilistic classifier, to explore feature subsets in tiny datasets. The assumption of feature independence is sometimes a good approximation. Especially effective for categories or text properties. The classification performance baseline provided by NB may assist researchers estimate feature selection method improvements. Thus, employing three classifiers, this research covered a wide

S. No	Datasets Used	Data type	Total Features	Instances	Class
1	SPECTF[72]	Integer, Real	44 + 1 (Class)	267	(0, 1)
2	PIMA[73]	Categorical	8 + 1 (Class)	768	(0, 1)
3	Breast Cancer[74]	Categorical	9 + 1 (Class)	286	(0, 1)
4	WBCD[75]	Real	31 + 1 (Class)	569	(0, 1)
5	WBCP[76]	Real	31 + 1 (Class)	198	(0, 1)
6	Liver Disorder[77]	Categorical, Integer, Real	6 + 1 (Class)	345	(0, 1)
7	ILPD[78]	Integer, Real	10 + 1 (Class)	583	(0, 1)
8	Hepatitis[79]	Categorical, Integer, Real	19 + 1 (Class)	155	(0, 1)

Table 3. Summary of datasets.

range of NCD data sets and feature selection methods to test the suggested algorithms on different circumstances and attribute kinds. Using NB, SVM, and DT classifiers, the algorithms will be tested for adaptation to varied data distributions and complexities. The system receives NCD data. Raw data may be absent or inconsistent, thus it must be processed before knowledge discovery. Figure 2 [81] shows classification.

Performance metrics

Numerous classification algorithms are available and continuously introduced. Assessing the effectiveness of these classifiers requires evaluating their performance. Various metrics are advocated as essential for accurately assessing the capabilities of an algorithm as a classifier[44].

Mathematically, Sensitivity is defined as

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

Specificity expressed as:

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

Precision computed as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Accuracy can be expressed mathematically as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

F1-score computed as:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

Kappa statistics quantify the agreement between observed accuracy and expected accuracy. This application is particularly valid in health and disease prediction, showing that the performance of the classifier is superior to random chance.

$$\mathrm{KS} = \frac{P_o - P_e}{1 - P_e} \tag{11}$$

Phase 1: result analysis

It guides H-GMRA in QRA for determining reduct set dependence value and cardinality. The suggested H-GMRA uses the QRA threshold gamma value and a 50-person population. H-GMRA and conventional QRA are compared on NCD datasets in this section. The comparison focuses on two areas. First, algorithms are assessed by reduct count and dependence ratio. Second, the best fitness values of both algorithms are compared.

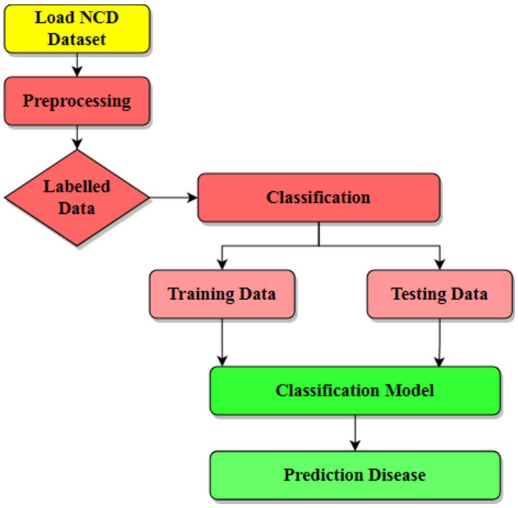


Fig. 2. [81] Classification process.

H-GMRA created a 21-attribute reduct with a dependence ratio of 0.8147 for the SPECTF heart dataset. H-GMRA's reduct set had different features than QRA, improving dependence accuracy by 3.57%. H-GMRA produced four reducts for the PIMA diabetic dataset, three of which had greater dependency ratios than QRA. The increased dependence accuracy values were 4.21, 4.08, and 3.69 higher than QRA. In the breast cancer dataset, H-GMRA produced two reducts with highest dependency ratios of 0.9683 and 0.9532 representing 2.22% and 0.71% increases. One of the WBCD dataset's two reducts had a maximum ratio of 0.8134, 1.48% greater than the QRA result. The WBCP dataset matched the QRA outcome, reducing two features. Extended computation time prevented us from presenting the ovarian cancer dataset's results. The liver dataset has five reduct attributes: two with dependence ratios of 0.9546 (a 3.74% improvement), 3.1% and 0.62%, and one with the same ratio but different characteristics. The ILPD dataset produced seven attributes with a dependency ratio of 0.9516, 2.89% greater than QRA. Finally, the hepatitis dataset yielded two reducts with a dependence ratio of 0.9565, a 2.12% improvement. Table 4 displays the reducts and dependence ratios for several NCD datasets, including dataset name, QRA and H-GMRA reducts, and their γQRA and γH-GMRA dependency ratios.

H-GMRA created a single reduct with 21 attributes, including F1R, F1S, F2R, F2S, and others, using the SPECTF heart dataset, resulting in an outstanding dependence ratio of 0.8147. Twelve of these 21 traits appear to be linked to stress. H-GMRA's reduct set has 3.57% higher dependence accuracy than QRA's. However, both algorithms share 11 properties, including F1R, F2R, F2S, and others. On the PIMA diabetes dataset, H-GMRA identified 7 attributes—Pregnancies, Glucose, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age—with a dependency ratio of 0.8724, surpassing QRA by 4.21%. Pregnancies, glucose, blood pressure, skin thickness, insulin, and pedigree function were shared by both methods. H-GMRA selected age, menopause, inv-nodes, deg-malig, and breast-quad for the breast cancer dataset, resulting in a maximum dependency ratio of 0.9683, a 2.22% improvement. QRA and H-GMRA shared tumor size and node-caps. H-GMRA found 6 features with a maximum ratio of 0.8134 for WBCD, 1.48% greater than QRA. Interestingly, both algorithms detected 3 traits. H-GMRA produced the same dependency results as QRA on the WBCP dataset, resulting in a reduct set of 2 features with distinct properties. First and second reducts shared 12 and 8 features, respectively. QRA findings for ovarian cancer are not shown here because to longer computation time.In the Liver dataset, H-GMRA found 2 reducts. The first reduct had mcv, alkphos, sgpt, gammagt, and beverages. The second reduct was same except sgot replaced sgpt. The maximum dependence ratio of these reducts was 0.9546, a 3.74% improvement. Along with gammagt and beverages, both reducts shared alkphos, sgpt, and sgot with the QRA reduct.On the ILPD dataset, H-GMRA reduced age, gender, DB, Alkphos, TP, ALB, and A/G to 0.9516, a greater dependence ratio. This ratio was 2.89% greater than QRA. Age, gender, alkphos, alb, and A/G were shared by both algorithms. H-GMRA reduced sex, anorexia, spiders, ascites, SGOT, and albumin for the hepatitis dataset, improving dependency ratio by 2.12%. Both algorithms recognized sex, anorexia, ascites, SGOT, and protime as common traits. All liver disease datasets showed alkphos, sgot, and sgpt. This emphasizes the importance of these three traits in identifying liver illness. Table 5 shows the best H-GMRA and QRA reducts for each NCD dataset. It shows the reduct with the largest threshold dependency from both techniques. Most datasets showed greater 'γ' values for H-GMRA compared to QRA, except for WBCP, where results were similar. The highest dependency values were highlighted.

S. No	Dataset	QRA		H-GMRA	
		Reducts	γQRA	Reducts	γH-GMRA
1	SPECTF	{8,5,1,4,11,6,3,10,18,12,7,27,31,20,2,4,32,28,30,36,40,41}	0.7790	{1,2,3,4,5,6,8,11,12,17,19,20,21,23,26, 27,30,35,36,39,42}	0.8147
2	Diabetes	{7,2,8,4,1,3,5}	0.8724	{1,2,4,5,6,7,8}	0.9145
				{1,2,3,4,5,6,7}	0.9132
				{1,2,3,4,6,7,8}	0.8621
				{1,2,3,4,5,7,8}	0.9093
3	Breast Cancer	{4,3,5,8,6}	0.9461	{2,3,4,7,9}	0.9683
				{3,4,5,7,8}	0.9532
4	WBCD	{1,3,14,22,24,27}	0.7986	{1,13,21,23,25,27}	0.8134
				{2,4,17,20,22,27}	0.7986
5	WBCP	{11,3,12,14,24,4,6,16,20,26,27,30,31}	0.8134	{3,4,5,11,12,14,16,20,24,26,27,30,31}	0.8134
				{3,4,7,11,12,15,16,20,24,26,29,30,31}	0.8134
6	Liver Disorder	{3,5,2,6,4}	0.9172	{1,2,3,5,6}	0.9546
				{1,2,4,5,6}	0.9482
				{1,2,3,4,6}	0.9546
				{1,3,4,5,6}	0.9234
				{2,3,4,5,6}	0.9172
7	ILPD	{3,8,1,10,5,2,9}	0.9177	{1,2,4,5,7,9,10}	0.9417
8	Hepatitis	{17,18,2,7,12,16}	0.9353	{2,7,11,12,16,17}	0.9565
				{1,2,7,10,15,18}	0.9518
				{2,7,10,16,17,18}	0.9353

Table 4. Feature Selection Results for QRA and H-GMRA on Various Datasets.

		QRA		H-GMRA	
S. No	Dataset	Reducts	γQRA	Reducts	γH-GMRA
1	SPECTF	{8,5,1,4,11,6,3,10,18,12,7,27,31,20,24,32,28,30,36,40,41}	0.7790	{1,2,3,4,5,6,8,11,12,17,19,20,21,23,26, 27,30,35,36,39,42}	0.8147
2	Diabetes	{7,2,8,4,1,3,5}	0.8724	{1,2,4,5,6,7,8}	0.9145
3	Breast Cancer	{4,3,5,8,6}	0.9461	{2,3,4,7,9}	0.9683
4	WBCD	{1,3,14,22,24,27}	0.7986	{1,13,21,23,25,27}	0.8134
5	WBCP	{11,3,12,14,24,4,6,16,20,26,27,30,31}	0.8134	{3,4,5,11,12,14,16,20,24,26,27,30,31}	0.8134
				{3,4,7,11,12,15,16,20,24,26,29,30,31}	0.8134
6	Liver Disorder	{3,5,2,6,4}	0.9172	{1,2,3,4,6}	0.9546
				{1,2,3,5,6}	0.9546
7	ILPD	{3,8,1,10,5,2,9}	0.9237	{1,2,4,5,7,9,10}	0.9516
8	Hepatitis	{17,18,2,7,12,16}	0.9353	{2,7,11,12,16,17}	0.9565

Table 5. Optimal reducts for QRA and H-GMRA on Various Datasets.

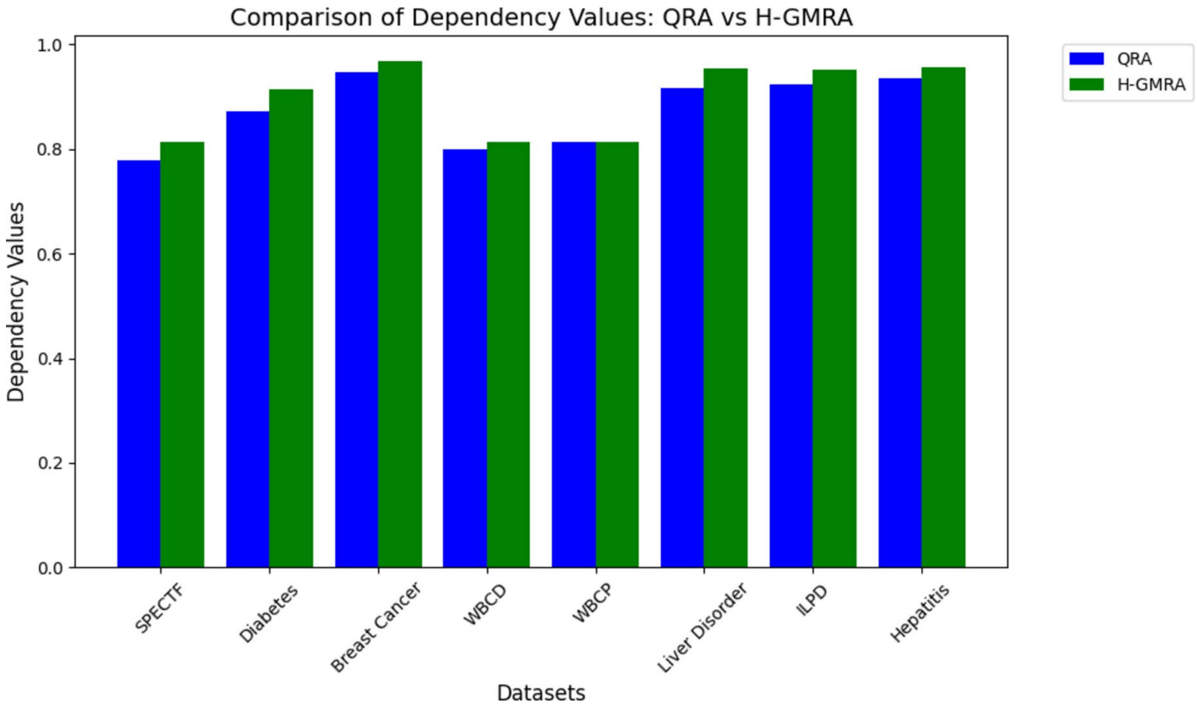


Fig. 3. Comparison of dependency values on NCD datasets.

Figure 3 shows that H-GMRA identifies smaller reducts with higher dependency ratios than QRA. Additionally, H-GMRA generates numerous reducts for each dataset while maintaining or improving dependence ratio. Figure 4 shows that H-GMRA routinely delivers reducts that outperform QRA. Certain datasets have identical H-GMRA and QRA dependency ratios but different feature selections. These findings show that the suggested approach can select features in NCD datasets. Multiple reducts may have overlapping properties, which might affect classification.

Phase 2: result analysis

MATLAB R2016a was used on a 2.60 GHz i5 processor, 64-bit Windows 8, and 4 GB RAM PC for research. The algorithm applied to eight NCD datasets. This section discusses algorithm performance and compares CPSO-RST-NFS to Quick_Reduct and H-GMRA.

The algorithm's hyperparameter setup ability is crucial for NCD dataset feature selection. The experiment hyperparameter settings are listed in Table 6. The settings were carefully chosen to balance feature selection and classification accuracy.

N determines the number of particles, while maximum iterations limits the optimization cycles. Particle exploration/exploitation is controlled by learning factors L1 and L2. The goal function's Gamma significance is controlled by custom parameters a and b. The optimization technique includes periodicity when adding any custom mathematical function f(x), including sin(x). To simplify trigonometric features, the argument angle, x,

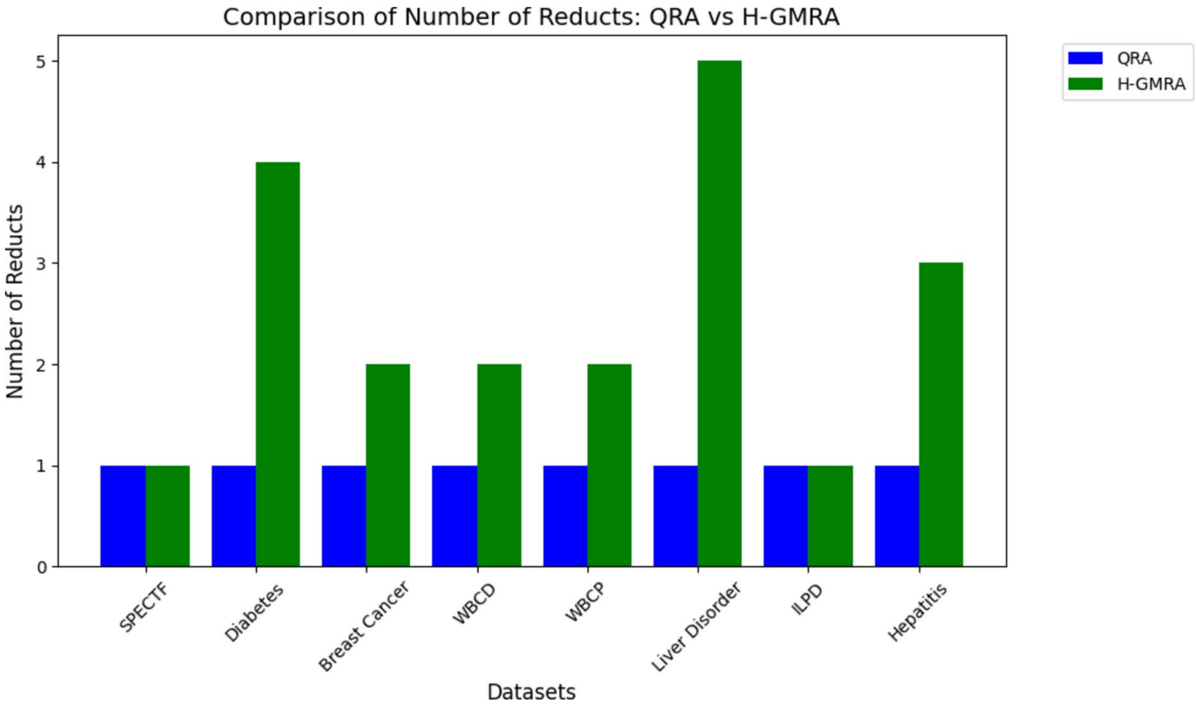


Fig. 4. Comparison based on number of reducts on NCD datasets.

Hyper parameter	Value/Setting
Population Size (N)	50
Maximum Iterations	100
Learning Factor (L1)	1.5
Learning Factor (L2)	1.5
Custom Parameter (a)	0.8
Custom Parameter (b)	0.2
Custom Function (f(x))	sin(x)
Angle Parameter (x)	π/4 (45 degrees in radians)

Table 6. Hyper parameter Settings for CPSO-RST-NFS Algorithm.

was adjusted to π/4, or 45 degrees in radians, as per mathematical use. We selected these hyper-parameters to best reflect NCD dataset feature selection settings while maintaining algorithmic clarity and efficiency. Iterative tests were used to balance feature relevance and computing efficiency during configuration. These settings can be adjusted to explore their effect on optimization outcomes, making CPSO-RST-NFS algorithm suitable for research under certain conditions.

The CPSO-RST-NFS algorithm is customized to maximize dependence value and minimize attribute set. This algorithm is always evaluated against QRA and H-GMRA. This approach is distinguished by its new goal function, which finds the optimal reduct with a high dependency value of 1 and a low feature count. Table 7 summarizes the performance of the CPSO-RST-NFS algorithm on non-communicable disease datasets.

Summary of the key findings is as follows:

- SPECTF Heart Disease Dataset: The algorithm achieved a best fitness value of 0.4675with a reduction of 72.73% in the number of features (from 44 to 12 attributes).
- Diabetes Dataset: It resulted in a 50% reduction in features with a best fitness value of 0.4254
- Breast Cancer Dataset: The algorithm obtained a best fitness value of 0.4487, reducing the feature set by 55.56%%.
- WBCD (Wisconsin Breast Cancer Database) Dataset: The algorithm reduced the attribute set by 70.97%, achieving a best fitness value of 0.4532.
- WBCP (Wisconsin Breast Cancer Prognostic) Dataset: This dataset had the best fitness value of 0.4678, with a substantial 70.97% reduction in features.
- Liver Disorder Dataset: The algorithm resulted in the best fitness value of 0.4182with only 5 selected attributes.

NCD Datasets	Total Features	No. of Selected Features	γCPSO-RST-NFS	Best Fit	Selected Features
SPECTF	44	12	1	0.4675	1, 2, 5, 9, 11, 13, 18, 19, 20, 30, 36, 42
Diabetes	8	4	1	0.4254	2, 4, 5, 6
Breast Cancer	9	4	1	0.4487	2, 3, 6, 8
WBCD	31	9	1	0.4532	2, 4, 9, 10, 12, 16, 20, 27, 30
WBCP	31	9	1	0.4678	1, 3, 6, 7, 12, 16, 17, 26, 31
Liver Disorder	6	5	1	0.4182	1, 3, 4, 5,6
ILPD	10	6	1	0.4393	1 ,2, 5, 6, 8, 9
Hepatitis	19	6	1	0.4665	1, 2, 3, 11, 14, 17

Table 7. Best Fitness Value of CPSO-RST-NFS Algorithm.

Dataset	QRA		H-GMRA		CPSO-RST-NFS	
	No. of Selected Features	γQRA	No. of Selected Features	γH-GMRA	No. of Selected Features	γCPSO-RST-NFS
SPECTF	21	0.7790	21	0.8147	12	1
Diabetes	7	0.8724	7	0.9145	4	1
Breast Cancer	**5**	0.9461	**5**	0.9683	4	1
WBCD	**6**	0.7986	**6**	0.8134	9	1
WBCP	13	0.8134	13	0.8134	9	1
Liver Disorder	**5**	0.9172	**5**	0.8134	5	1
ILPD	7	0.9237	7	0.9546	6	1
Hepatitis	6	0.9353	6	0.9546	6	1

Table 8. Comparison of QRA, H-GMRA and CPSO-RST-NFS.

- ILPD (Indian Liver Patient Dataset): It achieved a best fitness value of 0.4393, reducing the feature set by 40%.
- Hepatitis Dataset: The algorithm identified a subset of 6 features, representing a 68.4% reduction from the original set, with the best fitness value of 0.4665.

CPSO-RST-NFS reduced NCD dataset dimensionality while preserving or improving feature quality. The algorithm found the global optimum in feature selection by include the golden ratio, making it a promising method to dimensionality reduction in these data sets.

Comparative analysis of proposed CPSO-RST-NFS with QRA and H-GMRA methods — by number of selected features and by value of the dependency

Table 8 compares the proposed CPSO-RST-NFS technique to QRA and H-GMRA based on number of selected features and dependency value. This Table 8 shows how far the method under consideration is ahead of these two classification algorithms in feature selection and dependence value. This is necessary to evaluate CPSO-RST-NFS's feature set optimization with maximum relevance feature selection against the other two.

A comparison of the proposed CPSO-RST-NFS algorithm with QRA and H-GMRA in terms of the number of selected features and the dependency value makes it clear that this is effective when applied in feature selection.

- SPECTF Dataset: CPSO-RST-NFS achieved a 42.86% reduction in features compared to QRA and H-GMRA, while also improving the dependency accuracy by 28.43%.and 22.75% respectively.
- Diabetes Dataset: The proposed algorithm, CPSO-RST-NFS, selected 4 features, representing a 42.86% reduction in features and an 11.52% improvement in dependency ratio.
- Breast Cancer Dataset: CPSO-RST-NFS resulted in a minimal feature set similar in size to QRA and H GMRA with an additional 4.91% dependency parsing accuracy.
- WBCD Dataset: While CPSO-RST-NFS selected more features (9) than the other algorithms, it achieved a substantial 24.82% increase in dependency value.
- WBCP Dataset: CPSO-RST-NFS improved the dependency value by 22.93% compared to QRA and H-GMRA.
- Liver Disorder Dataset: This dataset saw an improved dependency value of 9.06% with the CPSO-RST-NFS algorithm.
- ILPD Dataset: For this dataset, CPSO-RST-NFS increased the dependency ratio by 8.12%.
- Hepatitis Dataset: CPSO-RST-NFS improved the dependency ratio by 7.29%.

Figure 5 shows CPSO-RST-NFS's feature selection compared to QRA and H-GMRA. Other than the WBCD dataset, CPSO-RST-NFS chooses less characteristics. Despite selecting additional features in WBCD, CPSO-RST-NFS has a high dependence ratio of 1. These findings demonstrate how well the proposed technique reduces feature set size and improves dependency values for various datasets. Figure 6 graphically compares dependency

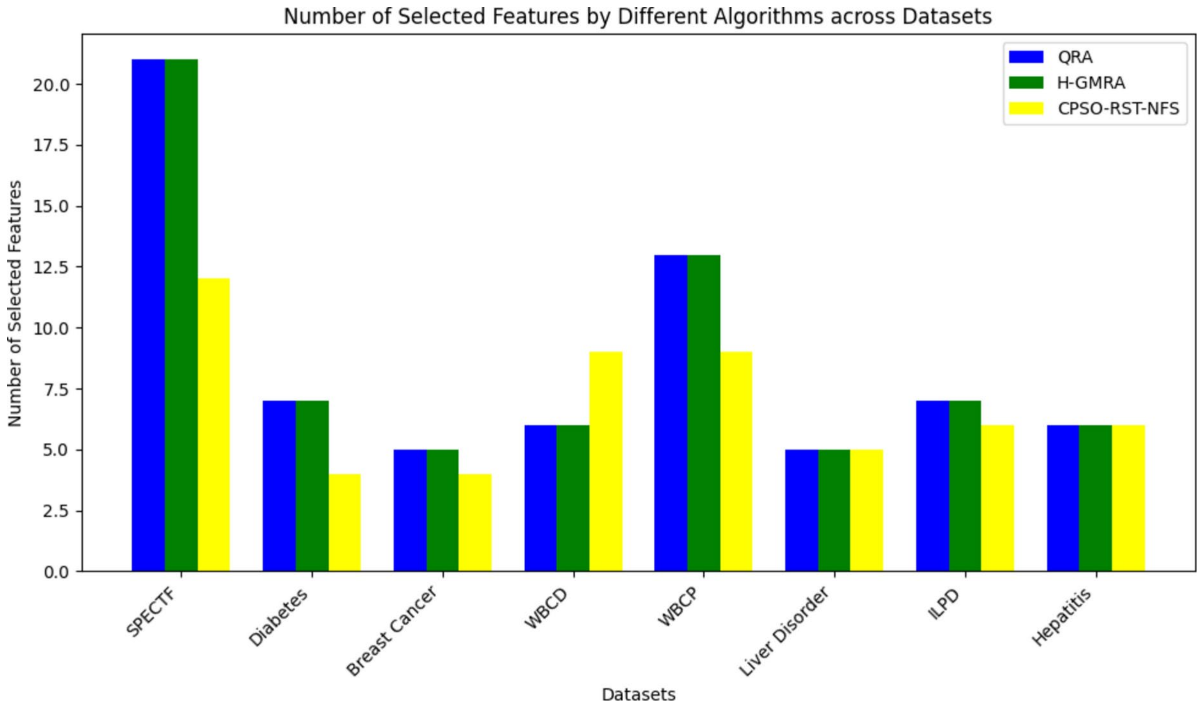


Fig. 5. Comparison based on number of features on NCD datasets.

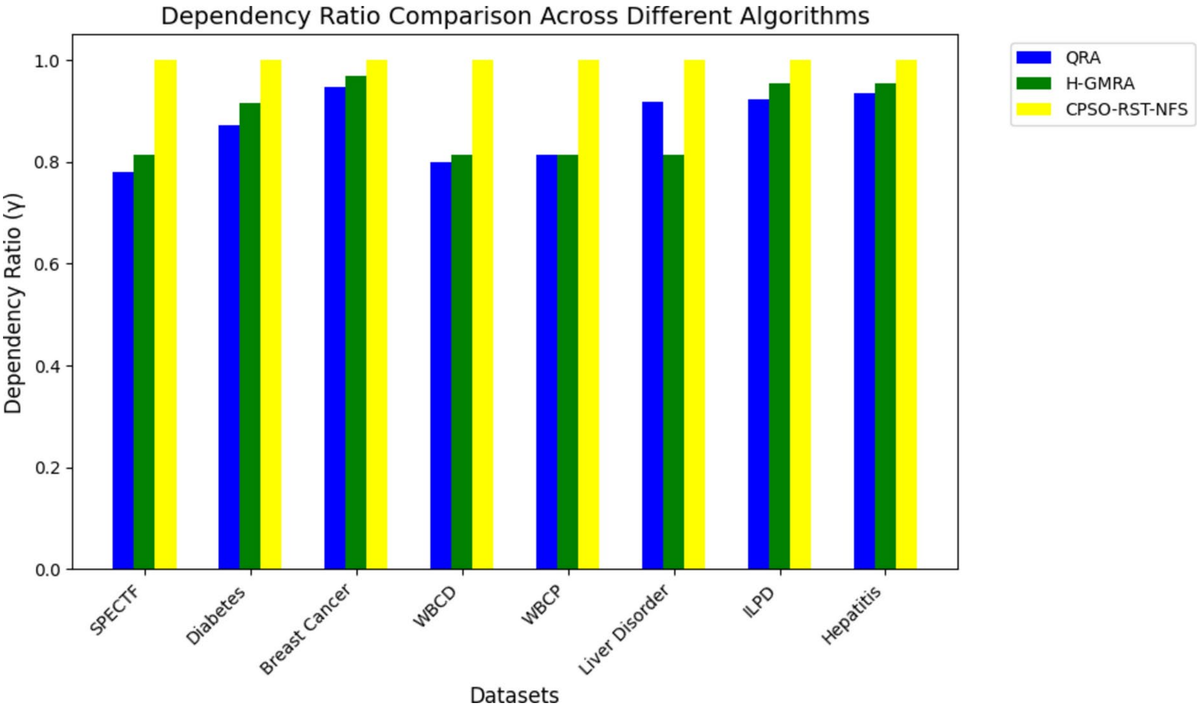


Fig. 6. Comparison based on dependency values on NCD datasets.

levels utilizing the proposed CPSO-RST-NFS algorithm, QRA, and H-GMRA. CPSO-RST-NFS outperformed the other two algorithms by assigning all datasets a maximum dependence value of 1. This graph shows that the CPSO-RST-NFS method can minimize features to maximize dependency and fitness for different NCD datasets. These findings demonstrate the algorithm's ability to minimize features while maintaining relevance, making it suitable for NCD dataset feature set optimization.

Performance analysis based on classifiers

The NCD dataset's performance on feature subsets from QRA, H-GMRA, and CPSO-RST-NFS algorithms will be evaluated using Naïve Bayes, Decision Trees, and SVM. All Weka experiments followed tenfold cross-validation, which divides datasets into training and test subsets. The algorithms were evaluated using classification accuracy, Sensitivity, Specificity, Precision, F-Score, Kappa Statistic, and Root Mean Squared Error.

Evaluation based on classification accuracy.

This section compares the classification accuracy of QRA, H-GMRA, and CPSO-RST-NFS algorithms utilizing support vector machines, decision trees, and Naïve Bayes. Each NCD dataset is evaluated similarly. Table 9 shows the analysis results. This investigation consistently shows that CPSO-RST-NFS classifies best in most NCD datasets. The analysis reveals that SVM is most accurate for five data sets, Naïve Bayes for four, and Decision Trees for three. This approach outperforms the others in accuracy when combined with the classifier in all datasets. The CPSO-RST-NFS technique is robust with many classifiers to ensure excellent classification accuracy across NCD datasets. Thus, it excels at feature selection and classification, with promising outcomes in practical applications.

Some classifiers do well on datasets. The CPSO-RST-NFS algorithms and Decision Trees achieved 87.3% accuracy on the SPECTF dataset. SVM accuracy was 77.1% for Diabetes and 73.8% for Breast Cancer. Decision Trees were most accurate on the WBCP dataset at 97.9%. The Naïve Bayes classification performed best on the Liver Disorder dataset, with an accuracy of 79.2%. SVM had 75% accuracy on the ILPD dataset in all three algorithms. On the Hepatitis dataset CPSO-RST-NFS achieved accuracy of 85% for SVM classifier. Figure 7 shows CPSO-RST-NFS-based classification findings.

CPSO-RST-NFS excels Accuracy Considering All Features, QRA, and H-GMRA across datasets. CPSO-RST-NFS had the greatest DTaccuracy on the SPECTF dataset at 87.3%, 3.7% better than H-GMRA DT and 15.8% better than DT with all features. The Diabetes dataset showed CPSO-RST-NFS with SVM accuracy of 86.7%, 9.6% and 10.1% better than H-GMRA and all features. In the Breast Cancer dataset, CPSO-RST-NFS had 77.1% accuracy using NB, 3.1% and 4.8% better than H-GMRA and all characteristics. CPSO-RST-NFS with SVM outperformed H-GMRA SVM by 1.9% and all features SVM by 3.1% on WBCD (Fig. 8). In WBCP, CPSO-RST-NFS with DT scored 97.9%, 1.9% higher than H-GMRA and 8.3% higher than all features DT (Fig. 9). In the Liver Disorder dataset, CPSO-RST-NFS with NB outperformed H-GMRA NB by 9.8% and all features NB by 23.8% with 79.2%. On the ILPD dataset, CPSO-RST-NFS with DT scored 72.4%, 4.6% and 6.6% better than H-GMRA and all features DT. Final results in the Hepatitis dataset showed CPSO-RST-NFS with NB at 85.7%, up 0.7% from H-GMRA NB and 2.4% from all features NB (Fig. 10). On all datasets and classifiers, CPSO-RST-NFS performed best. Despite initially lower accuracies, it performed well in Liver Disorder and SPECTF, as well as WBCD and WBCP. These results demonstrate the usefulness of the CPSO-RST-NFS algorithm in improving classification accuracy across NCD datasets.

The comparison of the algorithms based on other evaluative measures.

Classifier performance evaluation includes Sensitivity, Specificity, Precision, F1-Score, Kappa Statistics, and more beyond classification accuracy. Specificity shows real negatives, while Sensitivity shows true positives. Sensitivity and Specificity often trade off, with increasing one decreasing the other. These measurements are independent and depend on forecasts. RMSE, MAE, and ER are error-related indicators, hence lower values indicate better model performance. RMSE is useful for evaluating prediction model quality. Table 9 shows how selected features perform across NCD datasets using various classification techniques, revealing classifier usefulness beyond accuracy.

Figure 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig. 17 and Fig. 18 show that CPSO-RST-NFS classification algorithms have high sensitivity and specificity across all datasets. For three datasets (SPECTF, WBCD, WBCP) and two datasets (Diabetes and Breast Cancer), CPSO-RST-NFS A-based classifiers obtain 90% sensitivity and specificity, which is good for diagnostics. For all datasets, CPSO-RST-NFS has the highest specificity values compared to QRA and H-GMRA-based classifiers with a complete feature set. The CPSO-RST-NFS-based Naïve Bayes algorithm achieved the maximum sensitivity, specificity, accuracy, F1-Score, Kappa Statistics, and RMSE

Datasets	Accuracy Considering All the Features (in %)			Accuracy of QRA (in %)			Accuracy of H-GMRA (in %)			Accuracy of CPSO-RST-NFS (in %)		
	NB	DT	SVM	NB	DT	SVM	NB	DT	SVM	NB	DT	SVM
SPECTF	68.9	75.7	83.6	72.5	71	71.6	81.2	79.4	78.3	83.4	87.3	85.1
Diabetes	75.3	72.9	76.6	76.7	74.4	77.3	77.6	74.5	77.1	77.8	75.2	86.7
Breast Cancer	72.3	70.6	74.7	73.1	70.3	75.3	74	69.2	73.8	77.1	67.7	76.3
WBCD	91.6	92.4	94.2	93.6	94.2	95.4	92	94.3	91.7	92.8	94.5	97.3
WBCP	80.2	89.6	85.4	93.1	95.7	95.7	92.7	96	93.6	93	97.9	94.1
Liver Disorder	55.4	56.3	69.3	61.8	57.5	70.2	69.4	65.2	71.5	79.2	77	74
ILPD	55.9	65.8	69.4	57.6	69.3	72.8	55.1	67.8	72.4	55.9	72.4	75
Hepatitis	83.3	84	84.6	84.4	82.1	85	85.7	84	82.3	83.6	84.2	85

Table 9. Classification Accuracy for Different Algorithms on Various Datasets.

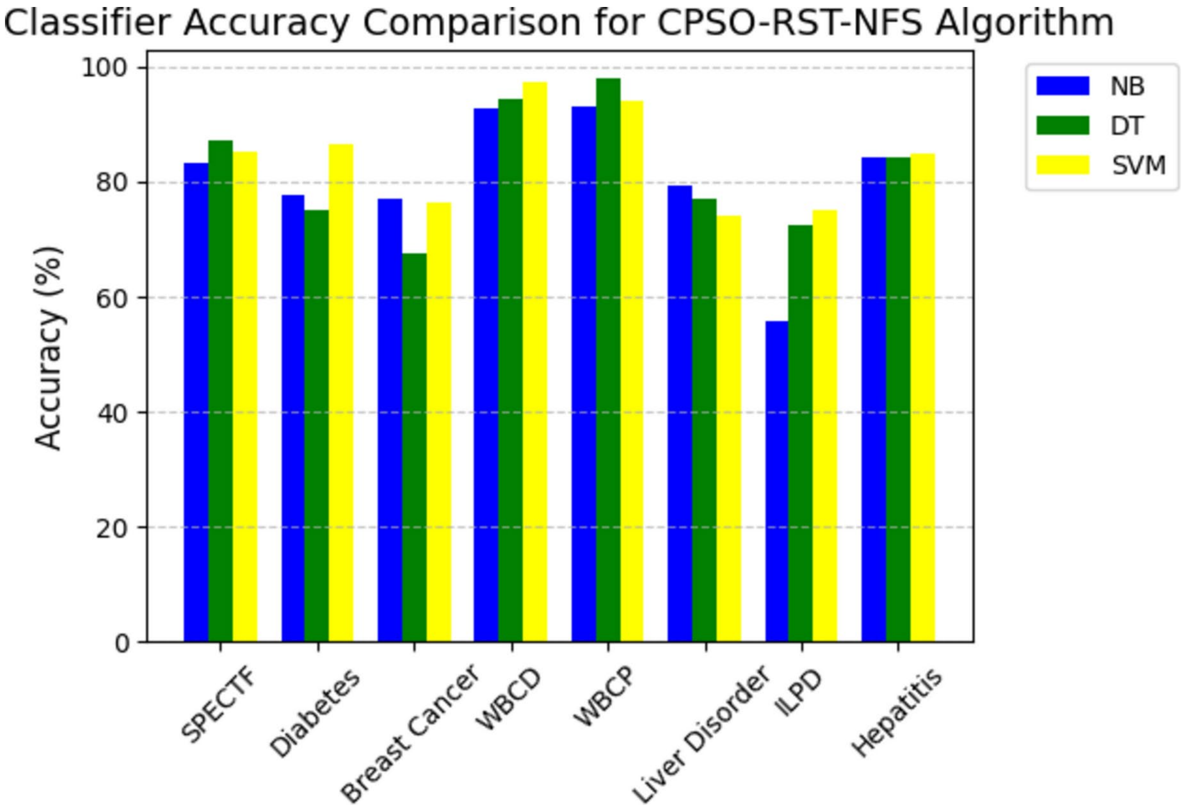


Fig. 7. Classification accuracy for CPSO-RST-NFS.

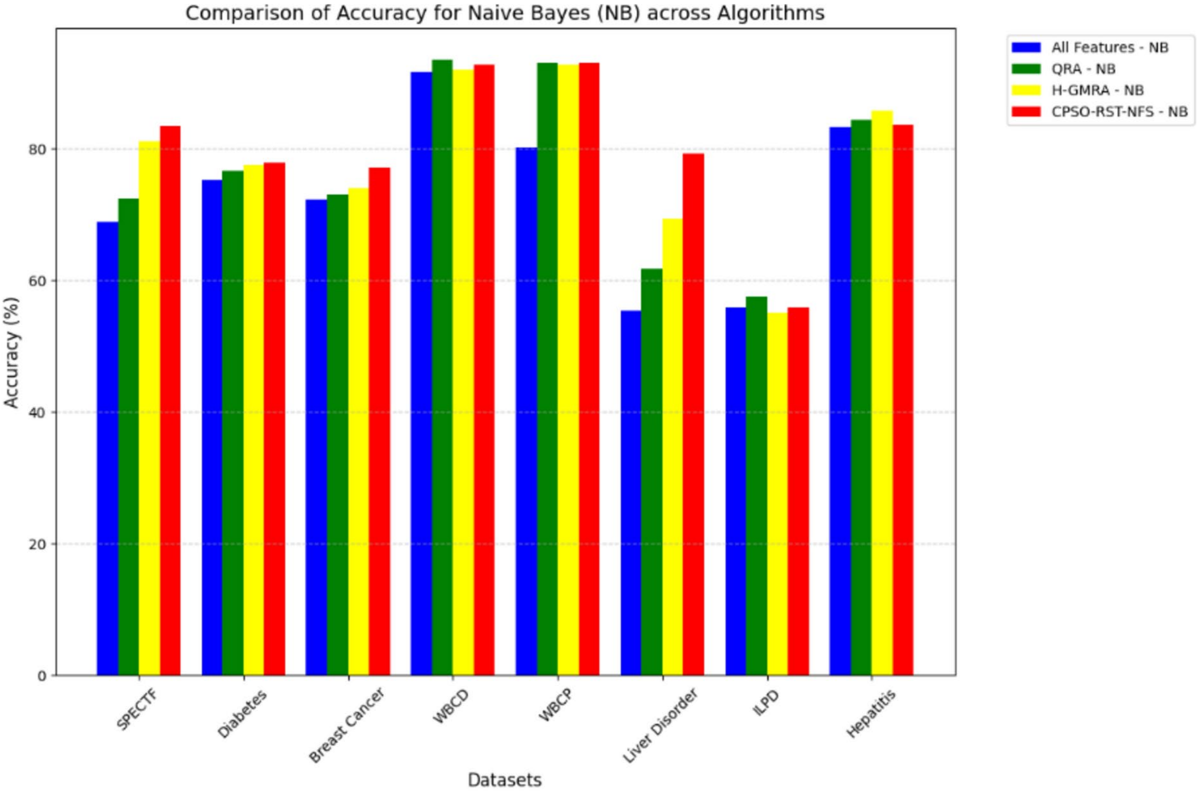


Fig. 8. Comparison of accuracy for NB across algorithms.

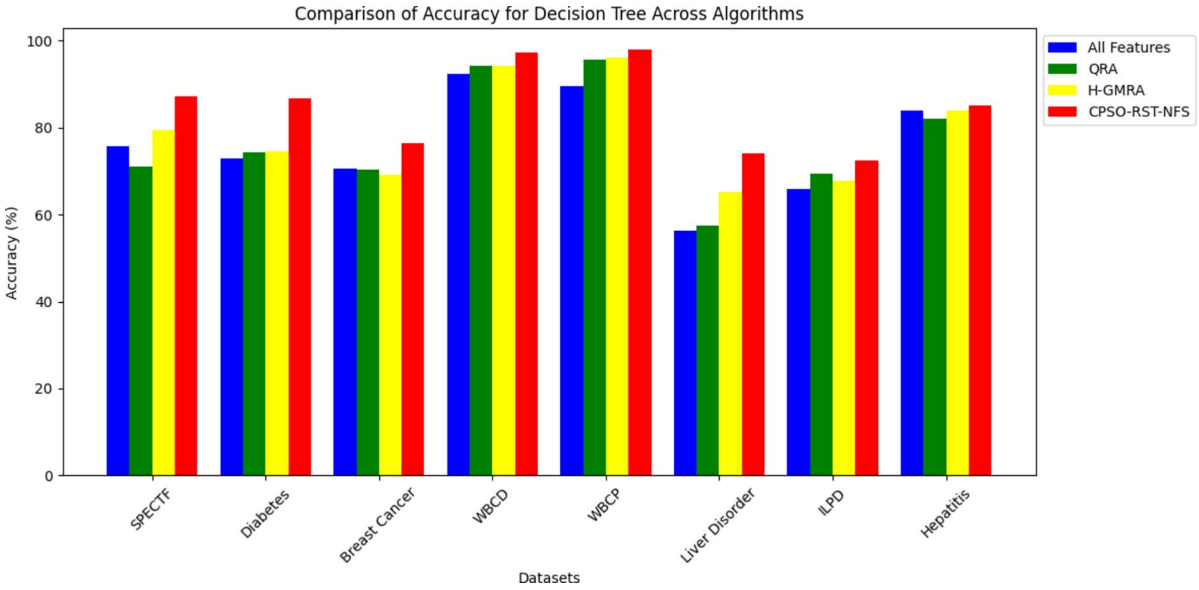


Fig. 9. Comparison of accuracy for DT across algorithms.

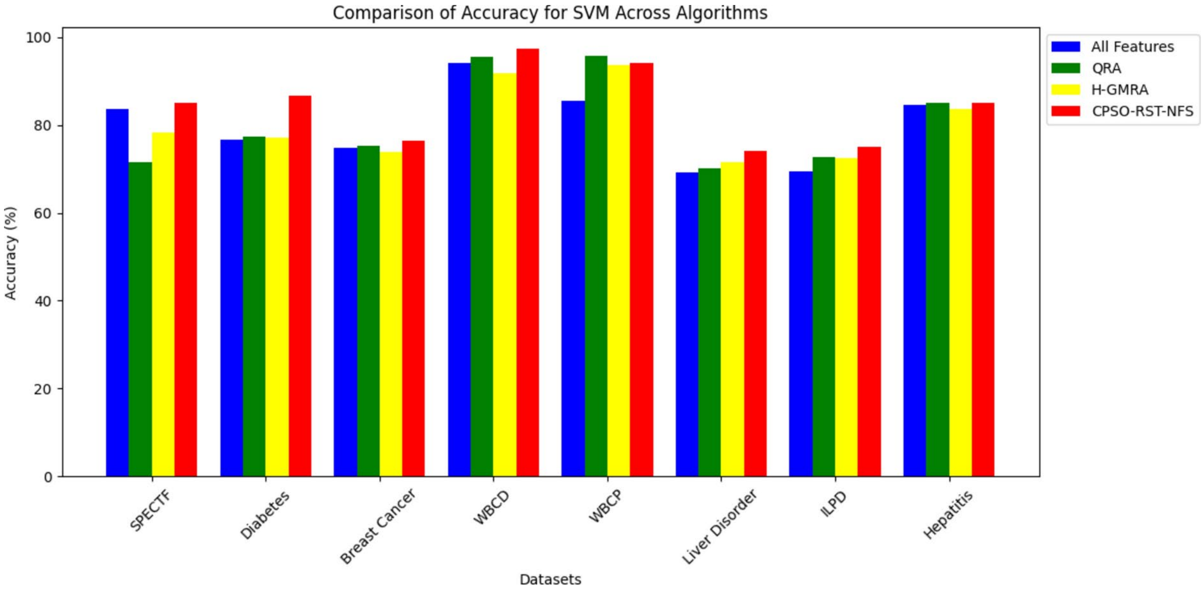


Fig. 10. Comparison of accuracy for SVM across algorithms.

for Liver Disorder and Hepatitis datasets. For the Diabetes dataset, it has the highest precision and lowest RMSE. For SPECTF and WBCP datasets, CPSO-RST-NFS-based Decision Trees DT method has the highest values for all evaluative measures, but for ILPD, it has the highest F1-Score. Finally, this CPSO-RST-NFS-based SVM method outperforms previous classifiers in Diabetes, Breast Cancer, and WBCD, ILPD, and Hepatitis datasets. It achieves maximum values for all evaluative parameters in the SPECTF and WBCP datasets and the greatest F1-Score in the ILPD dataset.

From diverse datasets, the graphic compares the average iterations of three feature selection algorithms—QRA, H-GMRA, and CPSO-RST-NFS. Figure 19 shows that CPSO-RST-NFS regularly takes less iterations than QRA and H-GMRA, proving its efficiency. The QRA method requires more iterations than H-GMRA, but the difference is small. Most efficient in terms of average iterations, CPSO-RST-NFS achieves the required results across all datasets with less resources.

Figure 20 shows that the CPSO-RST-NFS method excels QRA and H-GMRA in stability and dependability across datasets, with smaller spreads and fewer outliers suggesting low objective function value fluctuation. QRA and H-GMRA can have higher median values, especially in WBCD and WBCP datasets, but their broader

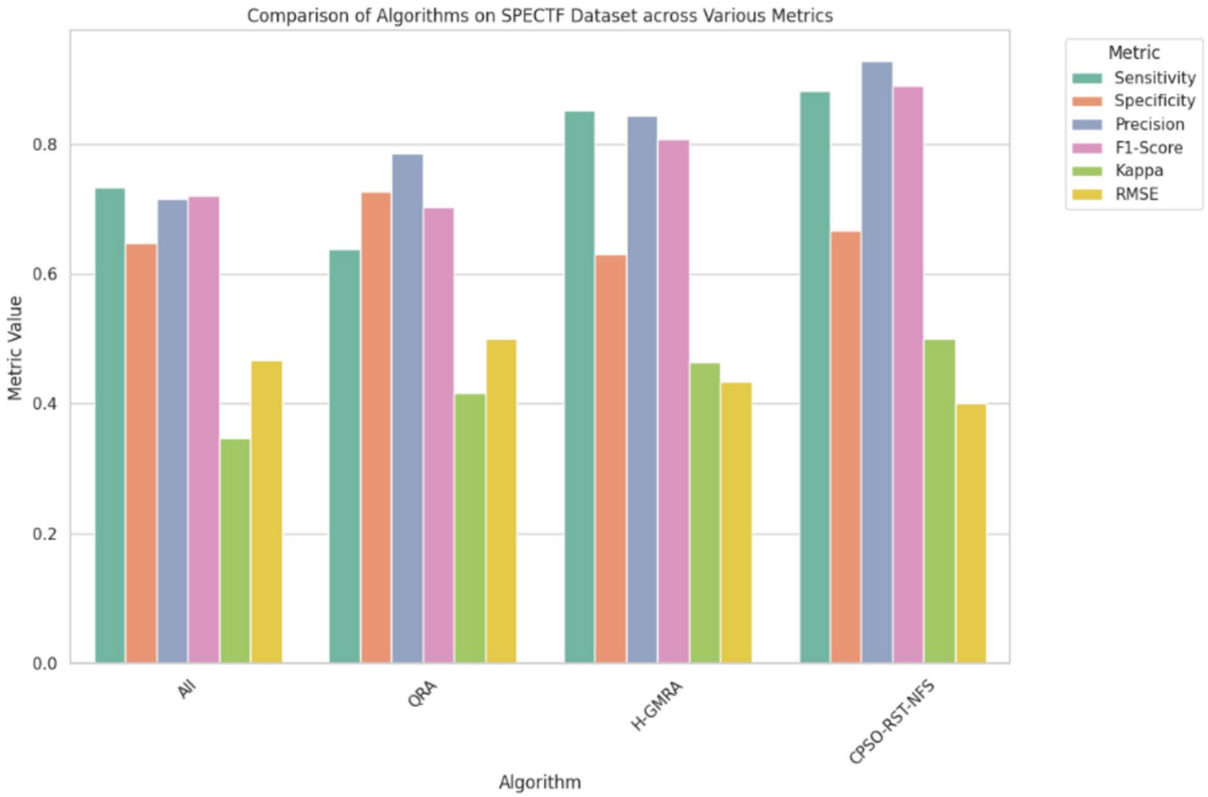


Fig. 11. Comparison of algorithms on SPECTF dataset for various metrics.

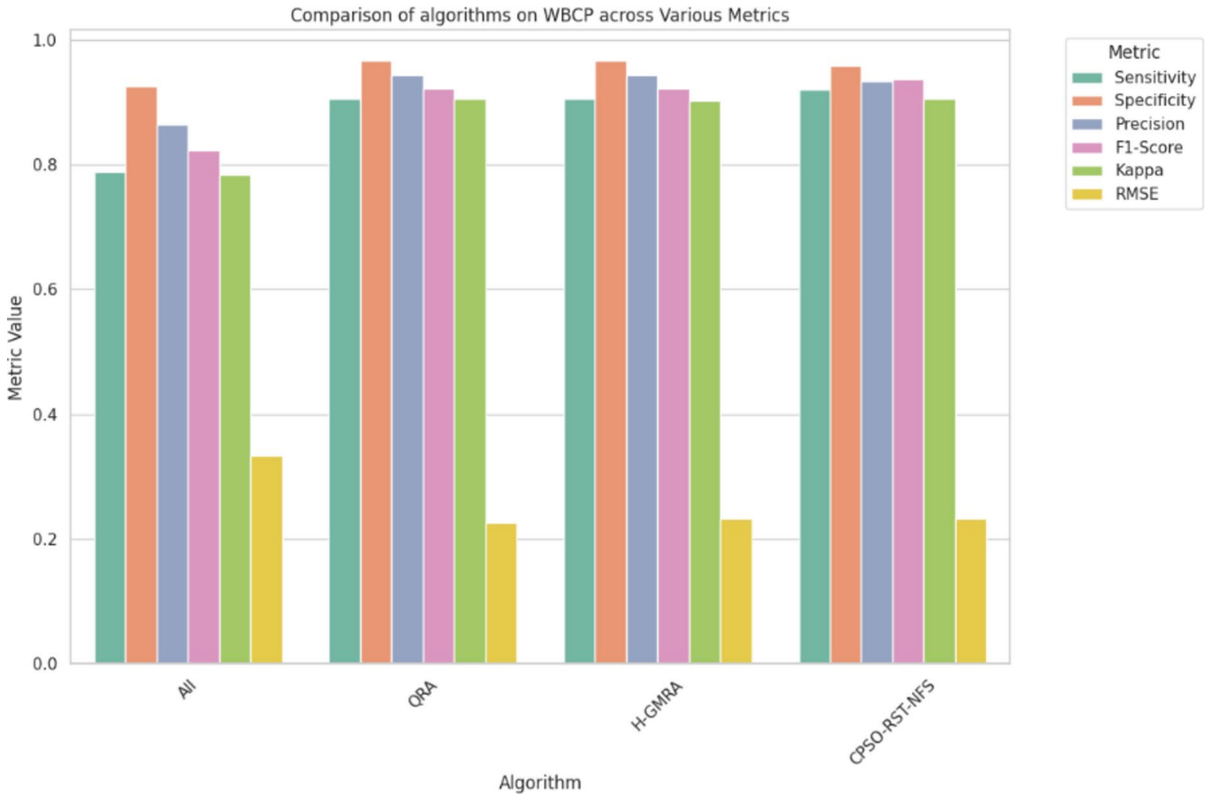


Fig. 12. Comparison of algorithms on WBCP dataset for various metrics.

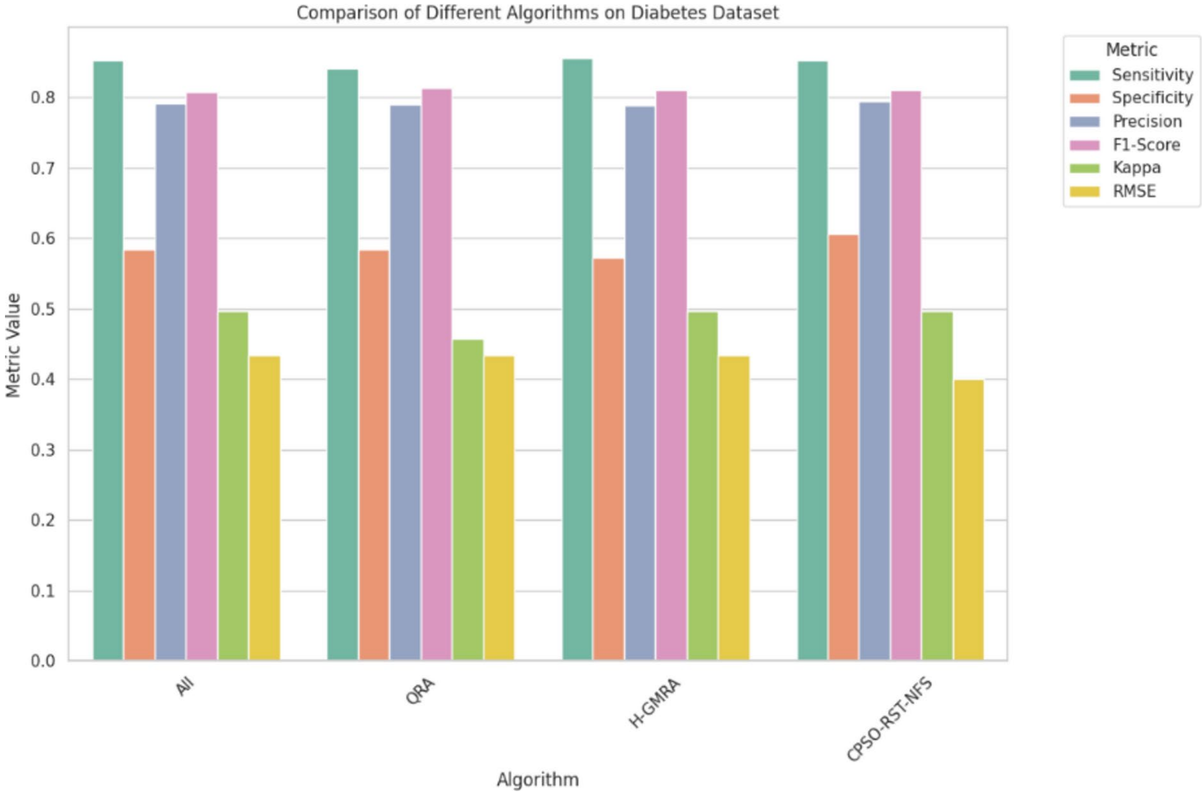


Fig. 13. Comparison of algorithms on Diabetes dataset for various metrics.

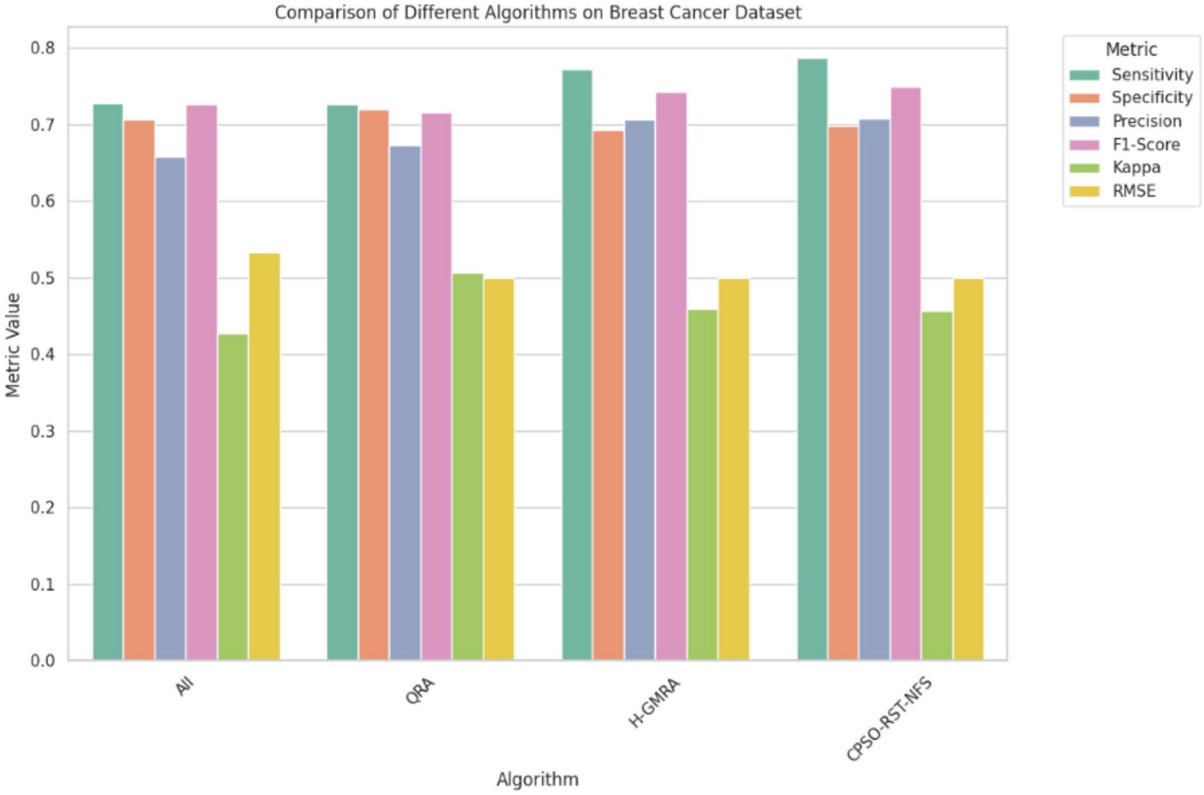


Fig. 14. Comparison of algorithms on Breast Cancer dataset for various metrics.

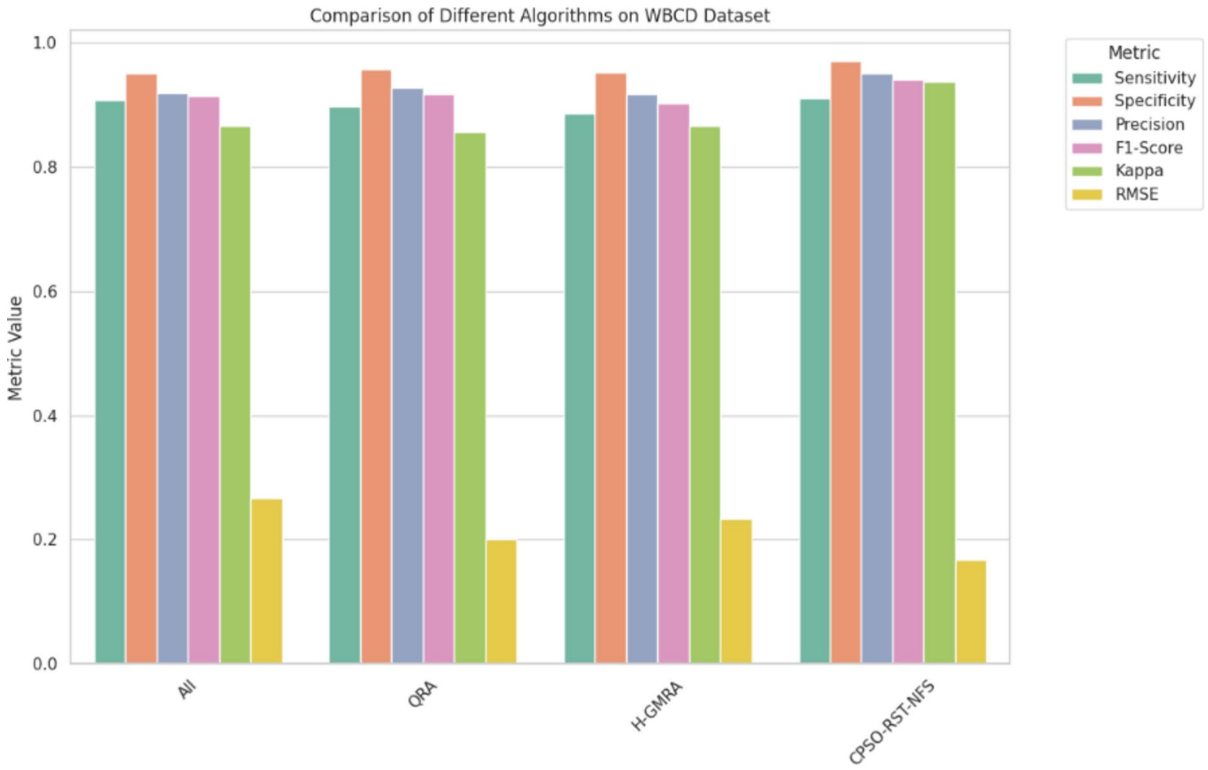


Fig. 15. Comparison of algorithms on WBCD dataset for various metrics.

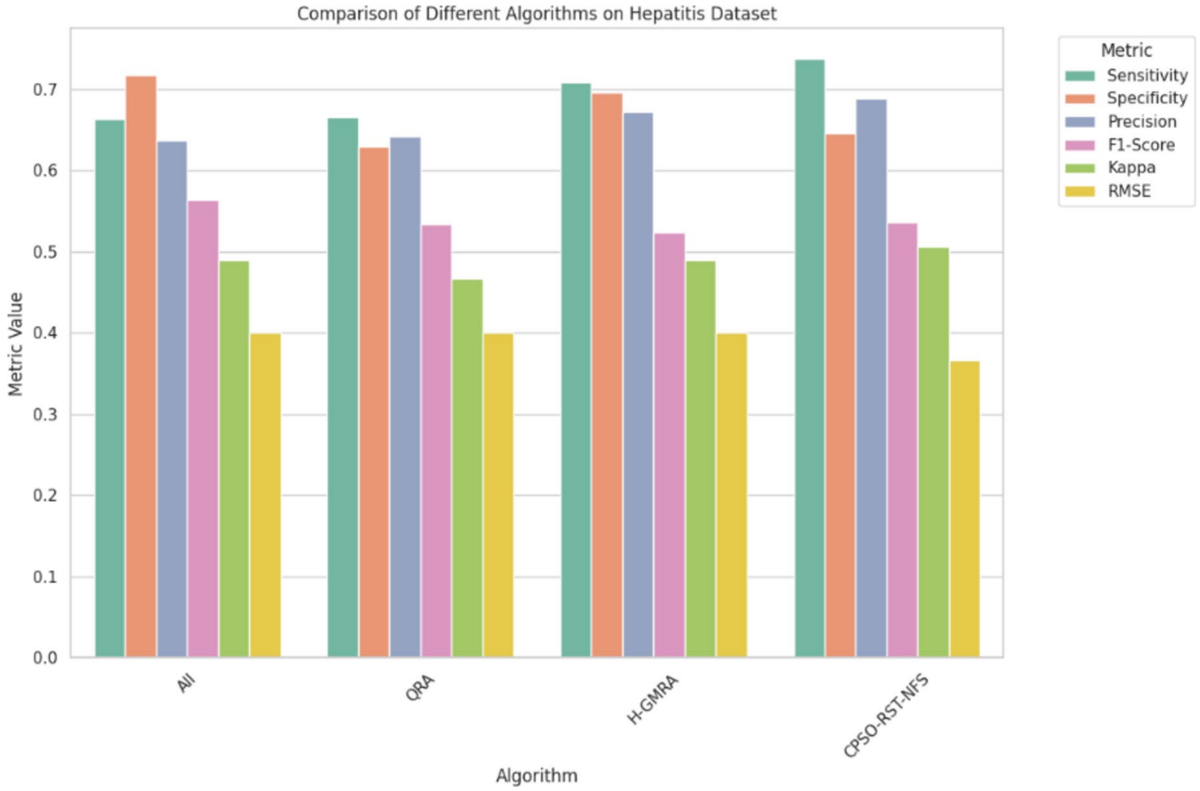


Fig. 16. Comparison of algorithms on Hepatitis dataset for various metrics.

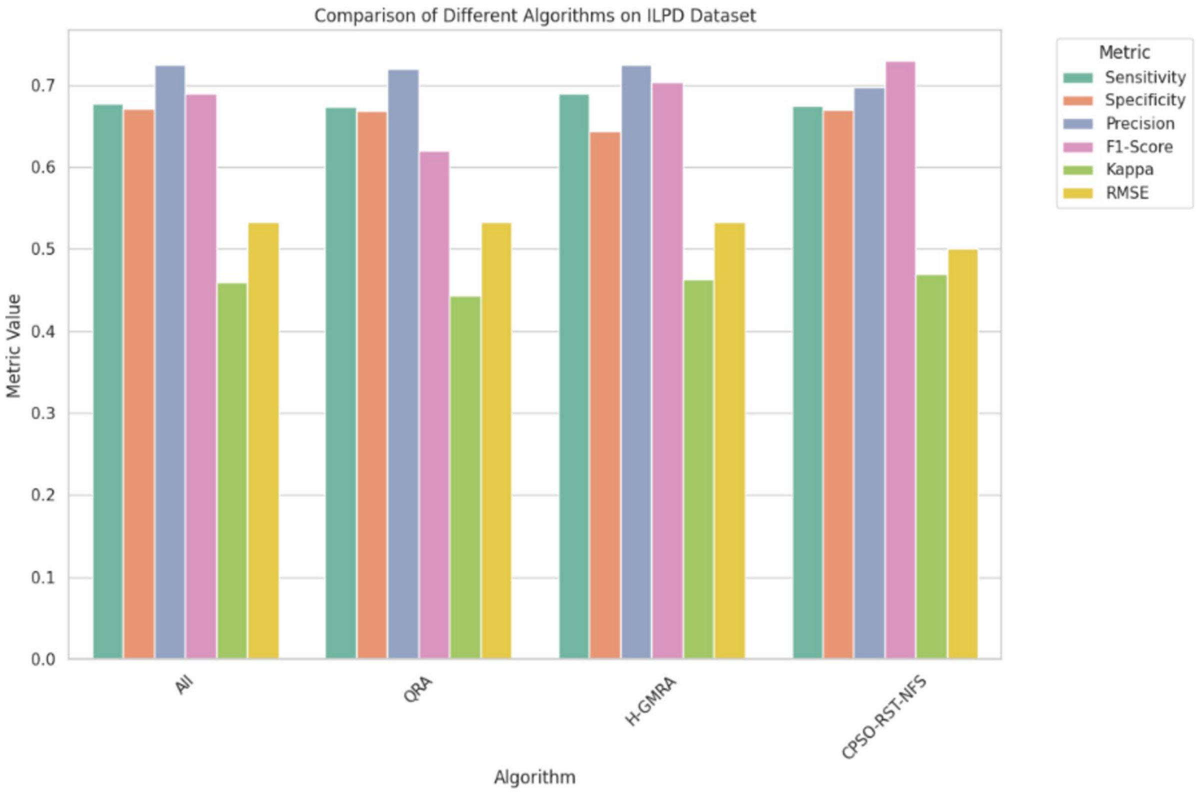


Fig. 17. Comparison of algorithms on ILPD dataset for various metrics.

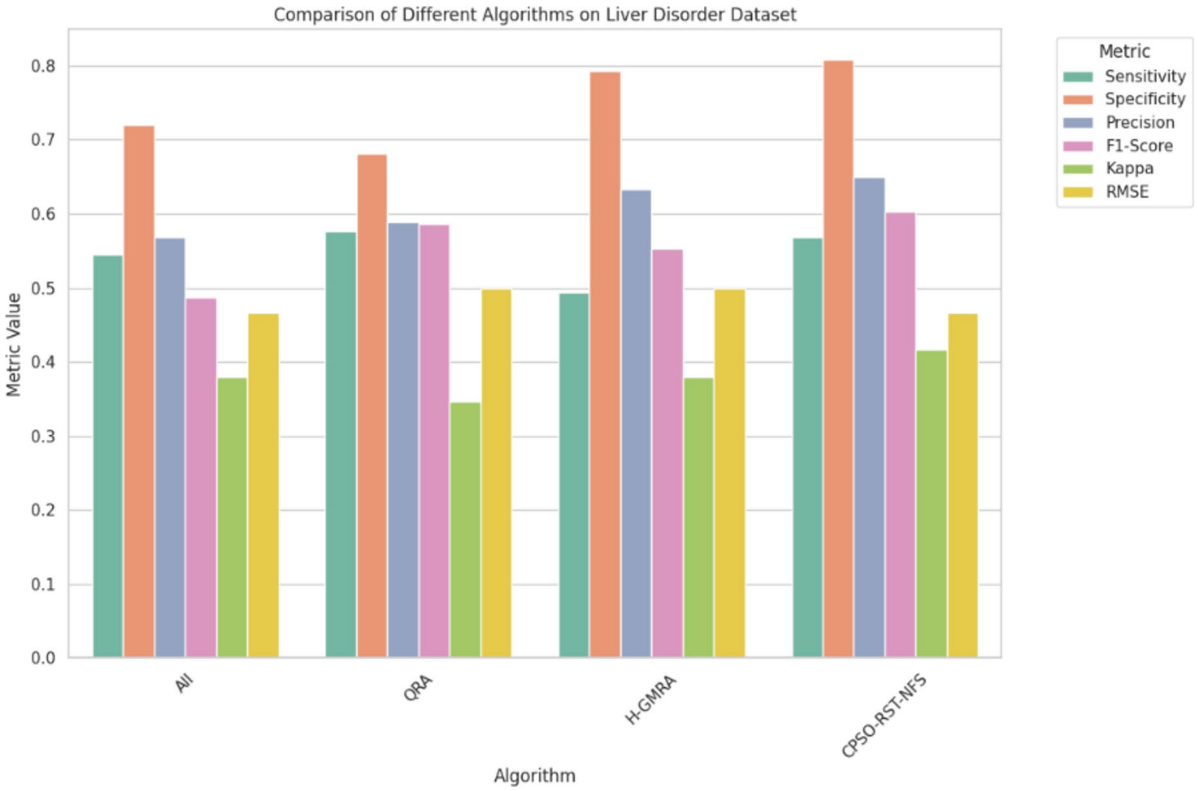


Fig. 18. Comparison of algorithms on Liver Disorder dataset for various metrics.

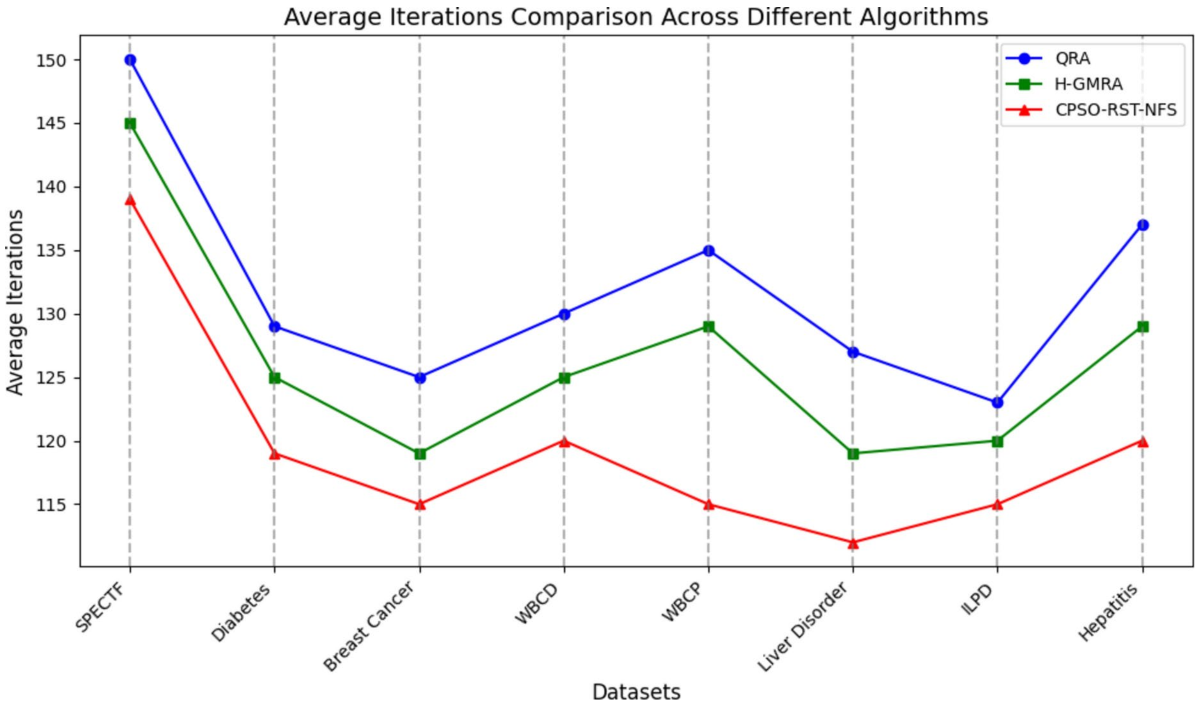


Fig. 19. Average iterations across datasets.

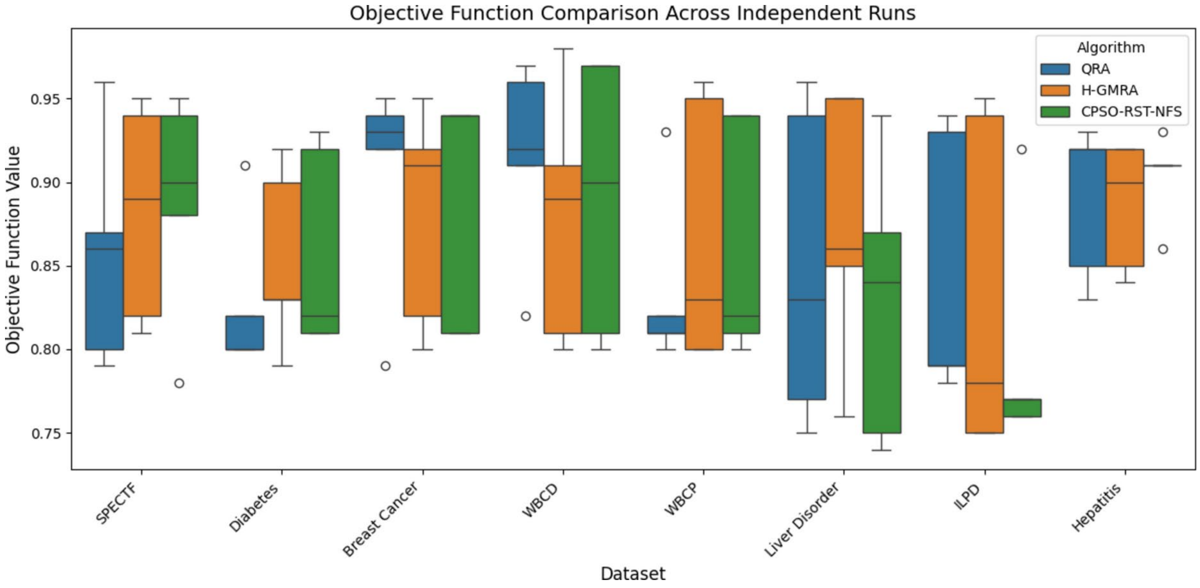


Fig. 20. Objective function comparison across different datasets.

spreads show substantial variability, which may affect their reliability. CPSO-RST-NFS performs consistently and has high median values in datasets like ILPD and Hepatitis.

Figure 21 shows QRA, H-GMRA, and CPSO-RST-NFS converged over 100 iterations on 8 datasets (SPECTF, Diabetes, Breast Cancer, WBCD, WBCP, Liver Disorder, ILPD, and Hepatitis). Three subplots display objective function values as algorithms advance for each dataset. Over time, all three methods decrease objective values, showing convergence to optimum solutions. QRA declines quicker in the early iterations before stabilizing at lower values. QRA declines faster than H-GMRA. CPSO-RST-NFS has a more slow convergence with modest variations than the other two algorithms, suggesting it may take longer to find an ideal solution. H-GMRA balances convergence speed and stability, making it excellent for controlled optimization. Although slower, CPSO-RST-NFS may perform better in complicated problems that need substantial solution space exploration to avoid local minima and find a superior global solution.

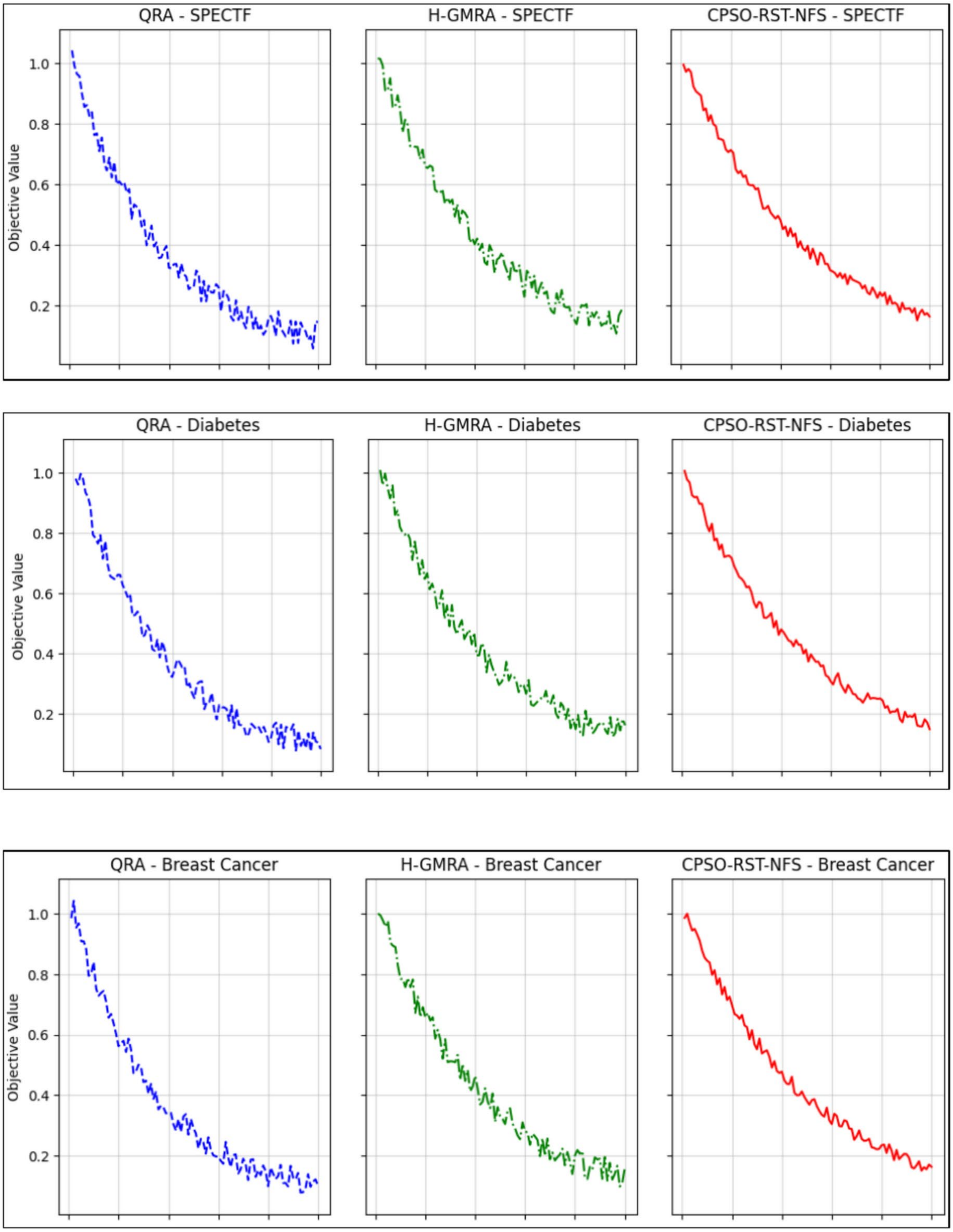


Fig. 21. Convergence comparison across different datasets.

Comparison to prior works

Table 10 shows Feature selection method comparisons on diverse datasets. The CPSO-RST-NFS approach yields good NCD results. On the SPECTF dataset, the CPSO-RST-NFS selects only 13 features with 88.7% accuracy, outperforming other approaches. Diabetes picks only 5 traits with 87.1% accuracy, exhibiting competitive performance. CPSO-RST-NFS selects 4 features with 97.9% accuracy on the WBCD dataset, demonstrating its feature selection effectiveness. With only 3 features, the ILPD Liver dataset has 71.4% accuracy, making it useful for illness diagnosis. In the Hepatitis dataset, CPSO-RST-NFS identifies 5 features with 84.5% accuracy. This

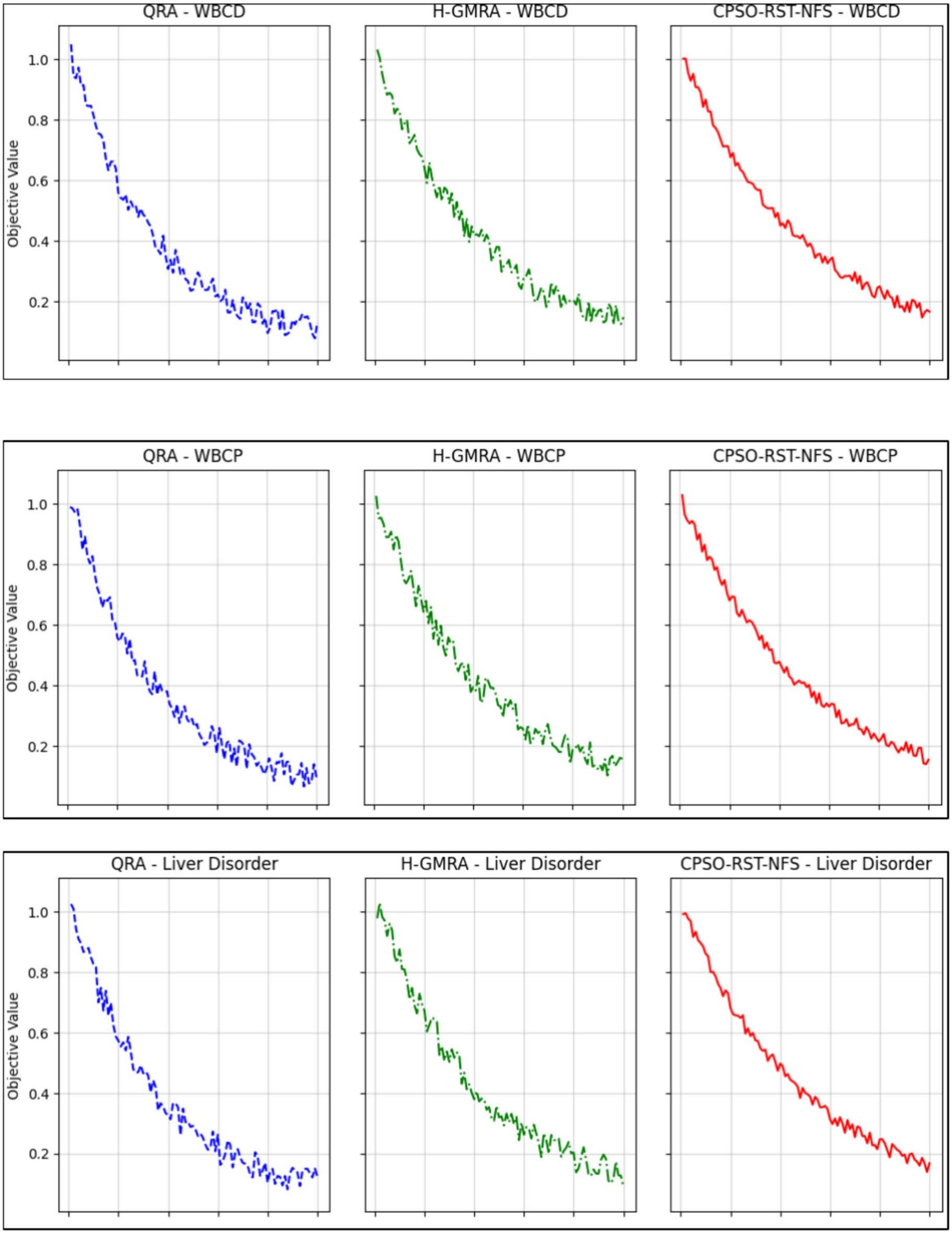


Figure 21. (continued)

suggests feature selection optimization. Overall, this technique is strong and effective because it balances feature selection and classification accuracy for NCD research and diagnostic applications.

Limitations and future work

The fundamental limitation of this work is focusing just on feature selection and classification accuracy. Still, it ignores other important NCD diagnosis factors as interpretability of selected characteristics and clinical relevance of the outcome. The authors' hybrid algorithms are intriguing, but their effectiveness varies between

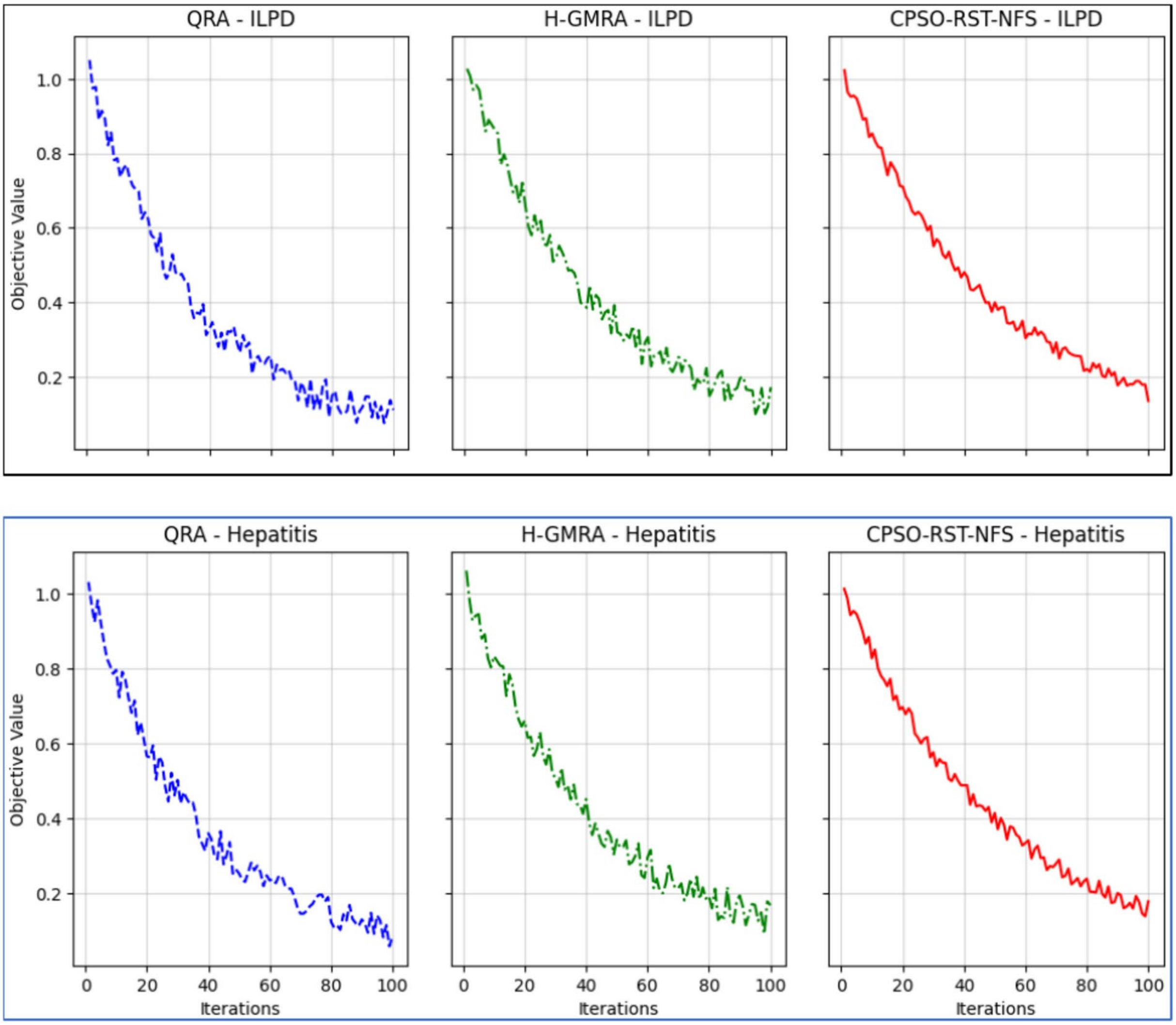


Figure 21. (continued)

NCD data sets, needing further fine-tuning and validation on more diverse real clinical data. Finally, all major datasets in this study have not accurately captured the complexity and diversity of real clinical scenarios, so further research on the validation of these findings in practical health care settings is needed. Future study should include interpretability and clinical relevance of features in addition to feature selection and classification accuracy.

This work proposes hybrid algorithms that need more testing across a larger number of real clinical datasets to improve robustness and generalizability. Additional data sources like electronic health records and patient demographics can be linked to improve NCD knowledge and prediction models. Finally, the algorithms will be implemented in easy-to-use healthcare software tools or platforms to improve NCD identification and management.

Conclusion

This study offers a solution to the challenge of diagnosis of NCDs through feature selection technique optimization. The due dates will improve the early detection of NCDs, and diagnosis always marks the beginning schedule of every treatment protocol. In this paper, two novel hybrid feature selection algorithms, namely H-GMRA and CPSO-RST-NFS, have been proposed and evaluated against different NCD datasets obtained from the UCI machine learning repository. First, research goes through detailed data preprocessing in handling missing values and normalization to assure the quality of datasets. The study then investigates the application of the new H-GMRA in contrast to traditional QRA. H-GMRA demonstrates its excellence by finding more than one reduce with high dependency value compared to QRA. It enhances the performance of NB, DT, and SVM classifiers. Even though SVM is the best classifier with further improvements in classifiers, when compared against all features, a problem of computation time is also addressed by this research, which includes Meta-heuristic algorithms. After the introduction of H-GMRA, the authors have implemented a filter-based algorithm of CPSO-RST-NFS. This technique performs feature selection with an optimized fitness function using a Golden

	Research papers			CPSO-RST-NFS	
Data Set	Methods adopted	Selected Features	Accuracy (%)	Selected Features	Accuracy (%)
SPECTF	Association Rules based Feature Selection [56]	14	77.14	12	87.3
	Ensemble Feature Selection [57]	19	86.5		
Diabetes	Ada Boost + Decision Stump [58]	-	80.72	4	86.7
	Improved NB [39]	-	82.3		
	Improved Electromagnetism-like mechanism [59]	-	77.21		
	Hierarchical and Progressive Combination of Classifiers [60]	-	83.34		
Breast Cancer	Ranker + SVM [61]	-	77.27	4	77.1
	SVM + CART [62]	5	73.03		
	Correlation Feature selection + Random Forest [38]	8	97.85		
WBCD	Threshold fuzzy entropy based feature selection[63]	31	97.28	9	97.3
	Hybrid SVM + RVM Classifier[64]	31	96.41		
	Non Linear Dualist Optimization Algorithm[41]	31	97.13		
	Hybrid PSO + SVM[65]	31	87		
	Hybrid ABC + SVM[65]	31	88		
WBCP	Hybrid SVM + RVM Classifier [64]	31	96.41	9	97.9
	Hybrid PSO + SVM [65]	31	88		
	Hybrid ABC + SVM[65]	31	87		
Liver Disorder	SVM [66]	6	70	5	79.2
	Integrated GA + Case Based Reasoning (CBR) model[67]	5	68.98		
ILPD	SVM Classifier[68]	8	73.2	6	75
	Variable – Neighbor Weighted Fuzzy KNN approach[50] Random	10	77.59		
	Under Sampling method + Stability Selection Method + Random Forest Classifier [47]	10	76.77		
	Backward Elimination + Linear SVM[69]	5	82.9		
Hepatitis	Brain Storm Optimization Algorithm[80]	10	97.16	6	85
	Threshold fuzzy entropy-based feature selection[61]	10	85.16		
	Correlation based ensemble feature Selection Algorithm[32]	16	93.90		
	Integrated GA + Case Based Reasoning (CBR) model[67]	–	94.19		

Table 10. Comparison of CPSO-RST-NFS results with the recent researches. The symbol '—' represents that the number of features have not been specified by the author

Ratio, which in turn will permit faster convergence and optimality in results. The reduced attribute sets obtained are further evaluated with the help of various measures apart from accuracy, such as sensitivity, specificity, precision, recall, F1-score, Kappa statistics. Generally, CPSO-RST-NFS obtained the best classification accuracy for the entire dataset, more so when combined with SVM. The ability of finding the best subsets of optimal features that may improve the accuracy of disease classification helped make a big difference in the diagnosis of NCDs. This hybrid FS approach, which includes H-GMRA and CPSO-RST-NFS, proves well to show potential for increasing the accuracy of diseases diagnosis. All in all, it gives valuable insights and provides tools to healthcare practitioners and data analysts in furthering the endeavor of early NCD detection and better care.

Data availability

The data that supports the findings of this study are available within the article.

Code availability

The code used in the study would be made available upon reasonable request to the corresponding author.

Received: 10 November 2024; Accepted: 18 February 2025
Published online: 06 March 2025

References

1. Sangaiah, I. & Kumar, A.. Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction. *Cluster Computing*. **22**. https://doi.org/10.1007/s10586-018-1702-5. (2019)
2. Abualigah, L. & Dulaimi, A. J. A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm. *Cluster Comput.* **24**, 2161–2176. https://doi.org/10.1007/s10586-021-03254-y (2021).
3. Madhusudhanan, B. et al. An hybrid metaheuristic approach for efficient feature selection. *Cluster Comput.* **22**(Suppl 6), 14541–14549. https://doi.org/10.1007/s10586-018-2337-2 (2019).
4. Joseph Manoj, R., Anto Praveena, M. D. & Vijayakumar, K. An ACO–ANN based feature selection algorithm for big data. *Cluster Comput.* **22**(Suppl 2), 3953–3960. https://doi.org/10.1007/s10586-018-2550-z (2019).

5. Vijaya, J. & Sivasankar, E. An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. *Cluster Comput.* **22**(Suppl 5), 10757–10768. https://doi.org/10.1007/s10586-017-1172-1 (2019).
6. Budhi, G. S., Chiong, R. & Dhakal, S. Multi-level particle swarm optimisation and its parallel version for parameter optimisation of ensemble models: a case of sentiment polarity prediction. *Cluster Comput.* **23**, 3371–3386. https://doi.org/10.1007/s10586-020-03093-3 (2020).
7. Das, A. & Sengupta, S. & Bhattacharyya, S. A group incremental feature selection for classification using rough set theory based genetic algorithm. *Applied Soft Computing.* **65**. https://doi.org/10.1016/j.asoc.2018.01.040 (2018).
8. Malar, B., Nadarajan, R. & GowriThangam, J. A hybrid isotonic separation training algorithm with correlation-based isotonic feature selection for binary classification. *Knowl. Inf. Syst.* **59**, 651–683. https://doi.org/10.1007/s10115-018-1226-6 (2019).
9. Kurman, S. & Kisan, S. An in-depth and contrasting survey of meta-heuristic approaches with classical feature selection techniques specific to cervical cancer. *Knowl. Inf. Syst.* **65**, 1881–1934. https://doi.org/10.1007/s10115-022-01825-y (2023).
10. Cerrada, M. et al. A hybrid heuristic algorithm for evolving models in simultaneous scenarios of classification and clustering. *Knowl. Inf. Syst.* **61**, 755–798. https://doi.org/10.1007/s10115-019-01336-3 (2019).
11. Lin, S. W. et al. Parameter determination and feature selection for back-propagation network by particle swarm optimization. *Knowl. Inf. Syst.* **21**, 249–266. https://doi.org/10.1007/s10115-009-0242-y (2009).
12. Malik, S. et al. Hybrid raven roosting intelligence framework for enhancing efficiency in data clustering. *Sci. Rep.* **14**, 20163. https://doi.org/10.1038/s41598-024-70489-1 (2024).
13. Malik, S. et al. MutaSwarmClus: enhancing data clustering efficiency with mutation-enhanced swarm algorithm. *Cluster Comput.*. **28**, 188. https://doi.org/10.1007/s10586-024-04822-8 (2025).
14. Challapalli, J. R. & Devarakonda, N. A novel approach for optimization of convolution neural network with hybrid particle swarm and grey wolf algorithm for classification of Indian classical dances. *Knowl. Inf. Syst.* **64**, 2411–2434. https://doi.org/10.1007/s10115-022-01707-3 (2022).
15. Reséndiz-Flores, E. O., Navarro-Acosta, J. A. & Hernández-Martínez, A. Optimal feature selection in industrial foam injection processes using hybrid binary particle swarm optimization and gravitational search algorithm in the Mahalanobis-Taguchi system. *Soft Comput.* **24**, 341–349. https://doi.org/10.1007/s00500-019-03911-w (2020).
16. Gauthama Raman, M. R. et al. A hybrid approach using rough set theory and hypergraph for feature selection on high-dimensional medical datasets. *Soft Comput.* **23**, 12655–12672. https://doi.org/10.1007/s00500-019-03818-6 (2019).
17. Mafarja, M. M. & Mirjalili, S. Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. *Soft Comput.* **23**, 6249–6265. https://doi.org/10.1007/s00500-018-3282-y (2019).
18. Hameed, S. S. et al. A comparative study of nature-inspired metaheuristic algorithms using a three-phase hybrid approach for gene selection and classification in high-dimensional cancer datasets. *Soft Comput.* **25**, 8683–8701. https://doi.org/10.1007/s00500-021-05726-0 (2021).
19. Meenachi, L. & Ramakrishnan, S. Differential evolution and ACO based global optimal feature selection with fuzzy rough set for cancer data classification. *Soft Comput.* **24**, 18463–18475. https://doi.org/10.1007/s00500-020-05070-9 (2020).
20. Ding, Y., Zhou, K. & Bi, W. Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer. *Soft Comput.* **24**, 11663–11672. https://doi.org/10.1007/s00500-019-04628-6 (2020).
21. Al-Dallal, A. & Al-Moosa, A. Prediction of Non-Communicable Diseases using class comparison data mining. *Adv. Sci. Technol. Eng. Syst. J.* **4**(5), 193–206 (2019).
22. Anitha, S. Heart disease prediction using data mining techniques. *J. Anal. Comput.* **8**(2), 48–55 (2019).
23. Arfiani, A. & Rustam, Z. Ovarian cancer data classification using bagging and random forest. *AIP Conf Proc.* https://doi.org/10.1063/1.5132473 (2019).
24. Kumar, S. & John, B. A novel gaussian based particle swarm optimization gravitational search algorithm for feature selection and classification. *Neural Comput. Applicat.* **33**(19), 12301–12315. https://doi.org/10.1007/s00521-021-05830-0 (2021).
25. Al-Mufadi, A & Al-Hagery, Mohammed. USING PREDICTION METHODS IN DATA MINING FOR DIABETES DIAGNOSIS (2014).
26. Atrey, K., Sharma, Y., Bodhey, N. & Singh, B. Breast Cancer Prediction Using Dominance-based Feature Filtering Approach: A Comparative Investigation in Machine Learning Archetype. *Brazilian Archives of Biology and Technology*. **62**. https://doi.org/10.1590/1678-4324-2019180486 (2019).
27. Subhadra, K. V. Neural network based intelligent system for predicting heart disease. *Int. J. Innov. Technol. Exp. Eng.* **8**(5), 484–487 (2019).
28. Azar, A. T. & Banu, N. Rough Set Based Ant-Lion Optimizer for Feature Selection. *Conf. Data Sci. Mach. Learning Appl. (CDMA)* https://doi.org/10.1109/CDMA47397.2020.00020 (2020).
29. Abbas, S. & Jalil, Z. BCD-WERT: a novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm. *PeerJ. Comput. Sci* https://doi.org/10.7717/peerj-cs.390 (2021).
30. Abiodun, E. A. A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. *Neural Comput & Applic*. (2021).
31. Acharya, S. (2021, May 14). https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383. Retrieved from https://towardsdatascience.com.
32. Alvarez-Alvarado, M.A.-C.-R. Three novel quantum-inspired swarm optimization algorithms using different bounded potential fields. *Sci. Rep.* **11**, 11655. https://doi.org/10.1038/s41598-021-90847-7 (2021).
33. Christo, V. R. E. Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network. *Comput. Math. Models Med.* https://doi.org/10.1155/2019/7398307 (2019).
34. Fan, S. K. S. & Chih-Hung, J. An Enhanced partial search to particle swarm optimization for unconstrained optimization. *Mathematics* **7**(4), 357. https://doi.org/10.3390/math7040357 (2019).
35. Ghosh, M. G. Binary Genetic Swarm Optimization: A A Combination of GA and PSO for Feature Selection. *J. Intell. Syst.* **29**(1), 1598–1610. https://doi.org/10.1515/jisys-2019-0062 (2019).
36. M. A. Rahman, R. C. Ovarian Cancer Classification Accuracy Analysis Using 15-Neuron Artificial Neural Networks Model. 2019 IEEE Student Conference on Research and Development (SCOReD) (pp. 33–38). Malaysia: IEEE. https://doi.org/10.1109/SCORED.2019.8896332(2019).
37. Patel, H. Medical data classification using HS, PSO, Hybrid PSO-HS based feature. https://www.academia.edu/36748274/Medical_Data_Classification_using_HS_PSO_Hybrid_PSO-HS (2019).
38. R. Dhanya, I. R. A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection. International Conference on Intelligent Computing and Control Systems (ICCS). Madurai, India: IEEE. 10.1109/ ICCS45141.2019.9065563 (2019).
39. Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data* **6**, 13. https://doi.org/10.1186/s40537-019-0175-6 (2019).
40. Tarawneh, M. Hybrid approach for heart disease prediction using data mining techniques. Acta Scientific Nutritional Health.(3) 147–151 (2019).
41. Vijayeeta, P. & Das, M. N. A novel approach for classification of malignant neoplasm using non-linear dualist optimization algorithm. *Int. J. Innov. Technol. Explor. Eng.* **8**(6), 583 (2019).

42. Chakraborty, S., & Shaikh, S. & Chakrabarti, A. & Ghosh, R. A hybrid quantum feature selection algorithm using a quantum inspired graph theoretic approach. *Appl. Intell.*. **50**. 1775–1793. https://doi.org/10.1007/s10489-019-01604-3 (2020).
43. Chelly Dagdia, Z. Z. A scalable and effective rough set theory-based approach for big data pre-processing. *Knowl. Inf. Syst.* **62**, 3321–3386. https://doi.org/10.1007/s10115-020-01467-y (2020).
44. Chen, R. C., Dewi, C. & Huang, S. W. Selecting critical features for data classification based on machine learning methods. *J Big Data* https://doi.org/10.1186/s40537-020-00327-4 (2020).
45. Chen, Y. & Chen, Y. Feature Subset Selection Based on Variable Precision Neighborhood Rough Sets. *Int. J. Comput. Intell. Syst.* **14**(1), 572–581. https://doi.org/10.2991/ijcis.d.210106.003 (2021).
46. Dumbaugh M, H. R. Public Health and Global Societies: A survey course in Global Health. Chicago: Public Health and Global Societies (PUBH 110). Retrieved from https://pubh110.digital.uic.edu: https://pubh110.digital.uic.edu/ section-2–12-non-communicable-diseases-aging/ (2021)
47. Fathi, M. N.-K. A machine learning approach based on SVM for classification of liver diseases. *Biomed. Eng. Appl. Basis Commun.* (2020).
48. Seyyedabbasi, A., Tareq Tareq, W. Z. & Bacanin, N. An effective hybrid metaheuristic algorithm for solving global optimization algorithms. *Multimed Tools Appl* **83**, 85103–85138. https://doi.org/10.1007/s11042-024-19437-9 (2024).
49. Felman, A. Everything you need to know about heart disease. *MedicalNewsToday*. (2021).
50. Kumar, P. & Thakur, R. S. Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach. *Multimed. Tools Appl.* https://doi.org/10.1007/s11042-019-07978-3 (2021).
51. Kumar, S. J. 4 26). A novel gaussian based particle swarm optimization gravitational search algorithm for feature selection and classification. *Neural Comput. Appl.* **33**, 12301–12315. https://doi.org/10.1007/s00521-021-05830-0 (2021).
52. Lancet, T. COVID-19: a new lens for non-communicable diseases. *Lancet* https://doi.org/10.1016/S0140-6736(20)31856-0 (2020).
53. Mehdi Alirezanejad, R. E. Heuristic filter feature selection methods for medical datasets. *Genomics* **112**(2), 1173–1181. https://doi.org/10.1016/j.ygeno.2019.07.002 (2020).
54. Farouk, R. M. & Mustafa, H. I. Breast Cancer Classification Based on Improved Rough Set Theory Feature Selection. *Filomat.* **34**(1), 19–34. https://doi.org/10.2298/FIL2001019F (2020).
55. Rahman MA, M. R. Artificial neural network with Taguchi method for robust classification model to improve classification accuracy of breast cancer. (344, Ed.) PeerJ Comput. Sci, 7, e344. https://doi.org/10.7717/peerj-cs. (2021).
56. Qu Y, F. Y.. Future selection algorithm based on Association Rules. J Phys Conf Ser. (2019).
57. Neumann, U. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData* https://doi.org/10.1186/s13040-016 (2016).
58. Veena Vijayan V, A. R.. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. *International Journal of Computer Applications*, **95**(17) (2014).
59. Wang, K. J. & Adrian, A. M. An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. *J. Biomed. Inform.* https://doi.org/10.1016/j.jbi.2015.02.001 (2015).
60. Kaur, H. HPCC: An ensembled framework for the prediction of the onset of diabetes. 4th International Conference on Signal Processing, Computing and Control (ISPCC). (pp. 216–222) (2017).
61. Pritom, A. I., Munshi, M. A. R., Sabab, S. A. & Shihab, S. Predicting breast cancer recurrence using effective classification and feature selection technique. 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 310–314. https://doi.org/10.1109/ICCITECHN.2016.7860215 (2016).
62. Lavanya, D. P. Analysis of feature selection with classfication: Breast cancer datasets. (2011).
63. Jaganathan, P. & Kuppuchamy, R. A threshold fuzzy entropy based feature selection for medical database clssification. *Comput. Biol. Med.* https://doi.org/10.1016/j.compbiomed.2013.10.016 (2013).
64. SK, Sheoran. Breast cancer classification using big data approach. Paripex Indian J Res, 401–403 (2018).
65. Utami, D. A. & Rustam, Z. Gene selection in cancer classification using hybrid method based on Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) feature selection and support vector machine. *AIP Conf Proc. DOI* **10**(1063/1), 5132474 (2019).
66. Hashem, E. M. & Mabrouk, M. S. A Study of support vector machine algorithm for liver disease diagnosis. *Am. J. Intell. Syst.* **4**(1), 9–14. https://doi.org/10.5923/j.ajis.20140401.02 (2014).
67. Singh, S. T. Impact of genetic optimization on the prediction performance of case-based reasoning algorithm in Liver Disease. *Int. J. Perform. Eng.* **13**(4), 348–361 (2017).
68. Mamdouh E, M. E. https://www.researchgate.net/publication/272356715_ A_Study_of_support_vector_machine_algorithm_for_liver_disease_diagnosis. Retrieved Feb 20, 2020, from https://www.researchgate.net: https://www.resear chgate.net/publication/272356715_A_Study_of_support_vector_machine_ algorithm_ for_liver_disease_diagnosis (2014).
69. Akyol, K. & Gültepe, Y. A study on liver disease diagnosis based on assessing the importance of attributes. *Int. J. Intell. Syst. Appl.* **9**(11), 1–9. https://doi.org/10.5815/ijisa.2017.11.01 (2017).
70. Elhoseny, M. B. Effective Features to Classify Ovarian Cancer Data in Internet of Medical Things. Preprints 2018. https://doi.org/10.20944/preprints 201809.0390.v1
71. Kanyongo, W. & Ezugwu, A. E. Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Informat. Med. Unlocked*, **38**, 2023, 101232. https://doi.org/10.1016/j.imu.2023.101232.
72. https://archive.ics.uci.edu/dataset/95/spect+heart
73. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
74. http://archive.ics.uci.edu/dataset/14/breast+cancer
75. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
76. http://archive.ics.uci.edu/dataset/16/breast+cancer+wisconsin+prognostic
77. http://archive.ics.uci.edu/dataset/60/liver+disorders
78. https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset
79. https://archive.ics.uci.edu/dataset/46/hepatitis
80. E, T. Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. Procedia Comput. Sci, (pp. 307–315). (2019).
81. http://hdl.handle.net/10603/379198

Acknowledgements

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/723/46.

Author contributions

Saleem Malik. S Gopal Krishna Patro and Chandrakanta Mahanty and K. Saravanapriya wrote the main manuscript text. Ayodele Lasisi. Quadri Noorulhasan Naveed prepared the figures. Anjanabhargavi Kulkarni helped code development. Abdulrajak Buradi. Addisu Frinjo Emma supervised the project. Naoufel Kraiem helped in review writing. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.M., S.K., A.F.E. or N.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.