



RESEARCH ARTICLE

OPEN ACCESS

IMPROVISED HARRIS HAWKS OPTIMIZATION WITH ROUGH SET MIN-MAX CLASSIFIER FOR FEATURE SELECTION TO ENHANCE THE DISEASE PREDICTION

Devi S^{*1}, V.S. Arulmurugan², V. Sujatha³, K.E. Narayana⁴, Saranya K S⁵,
M. Balamurugan⁶, J. Parthasarathy⁷

¹Department of Master of Computer Applications, Nandha Engineering College, Erode, India.

²Professor & Head, Department of Electrical and Electronics Engineering, Shree Venkateshwara Hi-Tech Engineering College (Autonomous), Gobichettipalayam, Erode, India.

³Professor, Department of Electronics and Communication Engineering, Shree Sathyam College of Engineering and Technology, Sankari, India.

⁴Assistant Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India.

⁵Associate Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathya Mangalam, India.

⁶Head of the Department & Associate Professor, Department of Computer Science and Engineering, The Kavery Engineering College, Salem, India

⁷Professor, Department of Information Technology, AVS Engineering College, Salem, India.

¹<https://orcid.org/0000-0002-7472-2319>, ²<https://orcid.org/0000-0002-2571-454X>, ³<https://orcid.org/0000-0009-4749-7353>,

⁴<https://orcid.org/0000-0002-4132-2175>, ⁵<https://orcid.org/0009-0003-3217-5817>,

⁶<https://orcid.org/0009-0004-0007-9406>, ⁷<https://orcid.org/0009-0006-7634-7045>

Email: *devikrishnamca@gmail.com, arulpvp@gmail.com, info@shreesathyam.edu.in, narayanake@gmail.com, saranyaks@bitsathy.ac.in, balacrazy.it@gmail.com, jpsphd@gmail.com

ARTICLE INFO

Article History

Received: January 23, 2026

Reviewed: March 24, 2026

Accepted: April 24, 2026

Published: May 29, 2026

Keywords:

Improved Harris Hawk
Optimization, Feature Selection,
Classification, Rough Set Min-
Max Classifier

ABSTRACT

The significant development in the medical field emerges a massive production of medical data, which must be appropriately analyzed to determine any disease. This sensitive information contains lots of redundancy or noise data which needs to be reduced through the Feature Selection (FS) approach. The researchers need an optimal system to solve this dimensionality reduction problem. Therefore, we proposed a novel Improved Harris Hawks Optimization (IHHO) algorithm based on Rough Set Min-Max Classifier (RMMC) to solve the FS problem using Chaotic-Improved Lévy Flight Distribution. IHHO is used to search for the best feature subset, whereas RMMC is used to design a clever rule for searching features using Euclidean distance measurement. Nearly three different datasets (Epileptic Seizure Recognition Data Set, Liver Dataset, and Breast Cancer dataset) are utilized to evaluate the performance of the proposed IHHO-RMMC model. According to our experimental results, the proposed IHHO-RMMC approach can solve FS problems and achieves 99.65% accuracy for Epileptic Seizure Recognition Data Set, 98.12% accuracy for Breast Cancer Dataset, and 96.22% accuracy for Liver Dataset which is higher than existing RMMC methods. The proposed IHHO-RMMC is expected to offer a high degree of accuracy in predicting diseases in the medical field.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Accurate diagnosis is important in determining effective therapy for a patient, and wrong diagnosis can result in technical peril for a physician. Manual examination of such a large medical dataset is not feasible for medical professionals and researchers. Methods in data mining simplify such an issue through extraction of meaningful patterns in a large dataset, and in return, reduce therapy costs. Feature selection (FS) is an important step in data mining, and it helps in choosing most useful features contributing towards disease prediction and classification. Filter, wrapper, and embedded approaches have been adopted for choosing an optimal feature subset, improving classification accuracy and reducing computational complexity.

Motivation: To build an Improved Harris Hawks Optimizer (IHHO) algorithm with integration of Rough Set (RS) theory for efficient feature selection, delivering sound medical diagnoses through minimizing unnecessary attributes, enhancing classification accuracy, and reducing computational cost.

Feature selection is significant for choosing important attributes, but traditional approaches have high computational costs and are susceptible to getting trapped in local optima. Nature-inspired search algorithms including Ant Colony Optimization and Grey Wolf Algorithm have boosted feature selection, with Harris Hawks Optimizer (HHO) offering even improved performance. Nevertheless, traditional HHO is plagued with repetitive search and getting trapped in local optima. To counter such vulnerabilities, in this work, HHO is upgraded with Improved Lévy Flight Distributions and chaotic mappings for effective feature subset selection for efficient disease prediction.

Problem Statement: Despite advances in FS techniques, present methodologies are still challenged with high-dimensional datasets, larger computation, and local optima issue. Rule-based feature classification algorithms reduce redundancy and identify dependencies between attributes, but conventional Rough Set (RS) theory is computationally costly. RS theory is effective in handling uncertain problems and uncovering hidden patterns by classifying objects into lower and upper approximations. Nevertheless, the integration of RS with optimization algorithms like IHHO (Improved Harris Hawks Optimization) requires further improvements to improve feature selection efficiency. The problem is to improve the Rough Set and IHHO architecture to optimize feature selection without affecting computational efficiency.

1.1 CONTRIBUTION OF THIS PAPER

- The main contribution of this paper is to solve the NP-hard [1] and classification accuracy problem through a novel IHHO-RMMC algorithm.
- Initially, the novel IHHO algorithm selects the best features from the disease data.
- Based on the selected features, RMMC rules are generated, and select the min-max rule to predict the disease.
- Then, the datasets are divided into training and testing data in the ratio of 70:30.
- The RMMC classifier classifies the disease class with the three datasets.
- To compare the performance of the novel IHHO-RMMC algorithm, the proposed model accuracy is compared with the existing model accuracy.

II. RELATED WORK

According to [2], RS is used to reduce redundant attributes using generated rules, and BP neural network filters all noise from data. This decision-making table avoids the model complexity in rules generation during neural network training, thereby improving the classification accuracy. The simulation results are compared with the actual reference value, whereas the model's data analysis still needs improvement in classification part. By [3] extracted massive amounts of medical data and analyzed using an intelligent model based on RST to generate a heart-diagnosing rule. The RST combined with the Boolean Reasoning Discretization algorithm to find all reducts. But this combination has more rules for extraction features that affects the model accuracy. In [4] investigated Rough Set Theory (RST) in dense and cs-dense sets for computing formulas to calculate the S-interior and S-closure operators. The SD-lower and upper is determined via these two sets. This paper completely analysis the RST behaviours in the boundary regions and its accuracy measure on sets. By [5] represented Binary Ant Lion Optimizer (BALO) is merged with two incremental hill-climbing approaches (QuickReduct and CEBARKCC) to develop HBALO model for solving the NP-hard problem. The filter model selects the intrusive features from the dataset for generating random solutions (ants). At last, the HBALO predicts the best result from the population, which is tested using the UCI repository. The HBALO mainly concern on NP-problem, it missed the improvement in searching best target value. According to [6] tried to address the FS issues using Brand-new Binary Whale Optimization Algorithm (BWOA) and RST.

Using 32 datasets from the UCI machine learning repository, the first experiment employed RST with BWOA, while the second used a wrapper method with BWOA on three distinct classifiers for feature selection. This BWOA cannot resolve higher-dimensional problems due to its low precision, sluggish convergence, and tendency to become stuck in locally optimal solutions. By [7] proposed a BPSOFPA, that combines Binary Particle Swarm Optimization (BPSO) and Flower Pollination Algorithm (FPA). PSO performs a global search in BPSOFPA, while the FPA search peculiarly to solve the binary problems and FS problems. Nearly two transfer functions are utilized, S-shaped and V-shaped, in the benchmark UCI dataset. The BPSOFA accuracy rate is lesser than the conventional method. According to [8], the GWNO is developed based on Grey Wolf Optimization (GWO) and Rough Set Theory (RST), where GWO support finding the minimal feature subset and RST provides a fitness value. The kNN classifier is utilized to validate the model performance using IHIS dataset. Due to the bad local search ability of GWO, the accuracy is affected. In turn [9] attempted to diagnose malignant leukemia early to provide the best treatment for the patient. In this paper, the author tried to analysis the acute lymphoblastic leukemia nuclei from the healthy lymphocytes using a backtrack search optimization algorithm. The feature extraction is carried out based on texture, color, wavelet, morphological, and statistical features.

Using Leukaemia dataset, the Jaya algorithm classifies the results whose performance is measured against NB, kNN, SVM, DT, and ensemble random undersampling boost. Jaya Algorithm has the disadvantage in dealing with an optimization problem. According to [10] proposed a new model with GWO and RST for mammogram images. Using mass-segmented mammogram images, the texture, shape features are extracted and generated as a feature set. Because of its performance limitation, the GWO-RST method needs to evaluate against the real-time health dataset. By [11], implemented a feature selection-based android malware detection using RST and Improved Particle Swarm Optimization (IPSO). The classic PSO algorithm encoding method is utilized to convert the data into the discrete domain, reducing the issues in maximum particle velocity and sigmoid functions. PSORS-FS ensures search process diversity and reduces the likelihood of premature convergence. The performance of this method was evaluated using UCI, the KEEL machine learning repository, and two Android permission datasets. In [12], analyzed brain tumor detection and classification using Rough set theory (RST) and Particle Swam Neural Optimization network (PSOENN). The MRI brain images are divided into train and test images. The PSOENN-RST model classifies the images into normal and abnormal. Still, the model needs to train with a higher epoch to improve the accuracy rate.

By analyzing the shortcomings of the conventional rough set theory, by [13] developed a new dimension-reduction strategy based on Binary Particle Swarm Optimization (BPSO) and probabilistic rough set theory. Two particular algorithms are part of the new methodology, PSOPRS, and PSOPRSN, which both add the number of attributes to the fitness function, while PSOPRS uses only the probabilistic rough set. When using Decision Trees (DT), Naive Bayes (NB), and Closest Neighbour Algorithms (CAN) to assess classification accuracy, PSOPRSN retrieves fewer attributes than PSOPRS while maintaining or slightly degrading classification performance. The model needs to reduce the attributes, but the attribute reduction affects the model performance. According to [14] researched Android malware with a feature selection technique based on an RST and an Improved Particle Swarm Optimization (IPSO) algorithm. The main contribution of this work is the recommendation of a new random key encoding method for converting the classical PSO approach to the discrete domain. It also reduces problems with maximum particle velocity and the sigmoid function related to binary PSO. The author was planning to implement the RS with the Artificial Bee Colony algorithm (ABC) to make the model to improve the FS robustly.

By [15] worked on a data mining task to Improve Harris Hawks optimization (IHHO), which selects the high-quality features for global optimization and provides faster convergence. The IHHO supports a better balance between exploration and exploitation. The IHHO working missed to concentrate on hawks energy calculation. In turn [16] proposed a Memory Energetic Harris Hawks Optimization (MEHHO) to select optimal features, memory saving, and learning strategy on feature selection with 13 low-dimensional and eight high-dimensional datasets. The learning strategy improves the learning ability of the hawks to reduce the local search. The MEHHO missed to calculate the hawk's prediction power. Based on the aforementioned analysis, the HHO is a good algorithm due to its derivative-free optimization technique. The main advantage of HHO is that it can be utilized for any optimization problems with the proper formulation. The conventional HHO aims to search the prey locally, but it needs an improvement in searching due to its repetitive search in the same region. So we developed an Improvised Harris Hawk algorithm (IHHO) with Chaotic-Improved Lévy Flight Distribution, which avoids the moving of hawks in the same region and supports it to enhance the searching behavior thereby providing an optimal solution for attacking prey. The rest of the paper is organized as follows: Sect.3 presents the proposed system. Sect.4 describes the experimental results. Finally, Sect. 5 concludes the research work and its future scope.

III. PROPOSED METHODOLOGY

Our paper proposes the IHHO-RST model, which combines an Improved Harris Hawks Optimization (IHHO) with a Rough Set Theory (RST) algorithm to increase the prediction accuracy and avoid searching in the same region using Chaotic-Improved Lévy Flight Distribution.

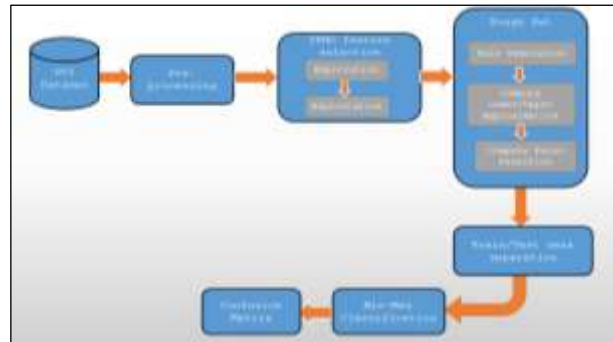


Figure 1: Architecture Diagram of the proposed HHO-RMMC algorithm.
Source: Authors, (2026).

III.1 IMPROVED HARRIS HAWK ALGORITHM

The development of the Conventional HHO algorithm is based on the inspiration of the hawks attacking the prey, which contains three processes: exploration, the transformation of exploration, and exploitation. In the Exploration stage, the prey location is randomly selected based on equation 1. let $P(t)$ represent the current position of the hawks, $P(t + 1)$ represent the hawk's new position for the next iteration, $P_{rabbit}(t)$ represent the prey position. Here r_1 to r_3 , q are random numbers generated in the interval of (0,1). UB and LB are the upper and lower bounds of variables. $P_{rand}(t)$ is the hawks random position in the location where $P_m(t)$ is the average location of the current population of hawks, r is the chance of successful prey escape. The hawks average position is calculated using equation (2). Let P_i is the position of each hawk in t iteration where N is the total number of hawks.

$$P(t + 1) = \{P_{rand}(t) - r_1|P_{rand}(t) - 2r_2P(t)|, (P_{rabbit}(t) - P_m(t)) - r_3(LB + r_4(UB - LB))\} \quad q \geq 0.5 \quad q < 0.5 \quad (1)$$

$$P_m(t) = \frac{1}{N} \sum_{i=1}^N P_i(t) \quad (2)$$

$$E = 2E_i \left(1 - \frac{t}{T}\right) \quad (3)$$

The hawks transition from exploration to the exploitation cycle where the energy of the prey is evaluated and represented as the equation (3) where E is the energy of the prey for escaping from hawks, E_i is the initial escaping energy which randomly changes in the interval of (-1,1) at each iteration.

When $|E| \geq 1$ hawk searching the prey in the exploration cycle, $|E| < 1$ finding a neighborhood of solution during the exploitation steps. The $|E| \geq 0.5, r \geq 0.5$, then the hawks in soft besiege and $|E| < 0.5, r \geq 0.5$, then the hawks in hard besiege.

$$P(t+1) = \Delta P(t) - E|P_{prey}(t) - P(t)| \quad (4)$$

$$P(t+1) = P_{prey}(t) - E|\Delta P(t)| \quad (5)$$

The soft besiege is represented as equation (4), the prey attempts to perform a surprise pounce where hawks try to encircle it softly. $\Delta P(t)$ is used to calculate the difference between the rabbit's current position for each iteration for escaping. In hard besiege, the hawks hardly encircle the prey which causes the prey to attain low escaping energy. The current position of prey is updated using equation (5). $|E| \geq 0.5, r < 0.5$ occurred during the soft besiege with progressive rapid dives where the prey have enough energy to escape with the surprise pounce which is called as leap frog movement or Levy Flight (LF) with zig zag deceptive motions. We proposed an Improved levy flight distribution with a chaotic algorithm for improving the speed of levy flight.

III.2 IMPROVED LÉVY FLIGHT DISTRIBUTION

The steady increment of the Lévy distribution follows this law. Lévy flight is characterized by many little steps and sporadically huge steps to avoid moving entities from repeatedly seeking the same region and changing a system's behaviour. Despite randomizing its motion direction, it exhibits an exponential rate distribution for its movement step size. The fitness function's starting value is the energy at each individual particle at the beginning of the initialization process, which initially executes all particles in the search space. The maximum energy is considered first, and the best iteration at that energy is considered the best particle sample. After that, the subsequent cycle is carried out per the equation for Levy flight as displayed in equation (6).

$$Improved\ Levy(x) = E_p + 0.01 \left(\frac{r_1 \delta}{|r_2|^{1/\varphi}} \right) \quad (6)$$

E_p is the Energy of prey. There are two random numbers between 0 and 1, r_1 and r_2 , φ which make up a constant of 1.5.

III.3 CHAOTIC MAPS

Known as chaotic maps, mathematical systems depict dynamic, deterministic processes highly dependent upon their initial conditions. Even if a predetermined procedure is followed, the consequences may be uncertain. Combined with metaheuristic algorithms, chaotic maps can enhance performance when applied to a particular optimization problem. Using them increases convergence and solves the local minima issue by getting closer to the best point by replacing random parts of the metaheuristic algorithm. Circle map chaotic equation (7) as

$$X_{k+1} = X_k + a - b/2\pi(\sin(2\pi X_k)) \quad (7)$$

The rabbit has enough energy to effectively escape when $|E| \geq 0.5$ but $r < 0.5$. At this point, a mild besiege is still being built before the surprise pounce. This process is smarter than the prior one. We assumed that the following rule in Eq. (8) would allow the hawks to assess (decide) their next action.

$$U = P_{rabbit}(t) - E|P_{rabbit}(t) - P(t)| \quad (8)$$

Using the following rule, we theorized that Hawks would dive depending on the Chaotic-Improved Levy Flight Distribution-based patterns.

$$V = U + (X_{k+1} + Improved\ Levy(x)) \quad (9)$$

As a result, Eq. (10) may conduct the final strategy for updating hawk positions during the soft besiege cycle.

$$P(t+1) = \{U \text{ if } F(U) < F(P(t)) \text{ W if } F(Z) < F(P(t))\} \quad (10)$$

Where U and V are obtained using Eqs. (8) and (9).

The rabbit cannot escape at $|E| < 0.5$ and $r < 0.5$, thus a hard besiege is built before the surprise pounce to capture and kill the victim. This step's prey side position is similar to that of the soft besiege, except this time, the hawks are attempting to close the gap between their usual location and the fleeing prey. Therefore, in a hard besiege situation, the following rule is followed:

$$P(t+1) = \{U \text{ if } F(U) < F(P(t)) \text{ W if } F(Z) < F(P(t))\} \quad (11)$$

$$U = P_{rabbit}(t) - E|P_{rabbit}(t) - P_m(t)| \quad (12)$$

$$V = U + (X_{k+1} + Improved\ Levy(x)) \quad (13)$$

Where $P_m(t)$ is obtained using Eq. (2)

Algorithm 1: IHHO Feature Selection**Inputs:** Population size N and Max. no Iterations T Initialization of parameter the P_i for N hawks**for** ($t = 1$ to T)Examine the fitness value of the hawks, $F(P)$ Set the best rabbit location as P_{rabbit} **for** ($i = 1$ to N)Calculate the initial energy E_i and jumping strength J as shown in

$$E_i = 2 \text{rand}() - 1, \quad J = 2(1 - \text{rand}())$$

Update E from (3)**if** ($|E| \geq 1$)

Update the location of the hawk using Eq. (1)

else if ($|E| < 1$)**if** ($r \geq 0.5$) and ($|E| \geq 0.5$)

Update the location of the hawk as shown in (4)

else if ($r \geq 0.5$) and ($|E| < 0.5$)

Update the location of the hawk solution in (5)

else if ($r < 0.5$) and ($|E| \geq 0.5$)

Update the location of the hawk solution in (10)

Using Chaotic-Improved Lévy Flight Distribution

else if ($r < 0.5$) and ($|E| < 0.5$)

Update the position of the hawk solution in (11)

Using Chaotic-Improved Levy Flight Distribution

end if**end if****return** i Update P_{rabbit} if a superior option exists**return** t **Output:** Global optimal best solution

The Improvised Harris Hawks algorithm (IHHO) is utilized to extract the best features from the disease data which is fetched to the RMMC algorithm and the result $A\{a_1, a_2\}$ is conditional attributes, and $D\{D_1, D_2\}$ is decision attributes which are represented in algorithm 1. Euclidean Distance [17] can be used to calculate the distance metric values for each attribute record. Calculate the neighborhood relationship with $\theta(x_1)$ by setting the $\theta = 0.1$. Depending upon the θ value separated the neighbourhood relation for attributes. The Euclidean formula is implemented in equation (14) in order to improve performance by evaluating neighborhood values

$$F(x_i, x_j) = \sum_{k=1}^n \sqrt{((x_i(k) - x_j(k))^2)} \quad (14)$$

A Euclidean distance is calculated for the selected features. This is satisfied by the new Euclidean distance formula, $\theta(x_1) = \{x_1, x_2\}$. The medical dataset is obtained from the University of California, Irvine (UCI) repository. Initially, normalization is used to convert negative values to positive values. The attributes' Neighborhood and Equivalence relations are calculated for the selected features. The conditional attributes granule values are then calculated with the "and" operator.

The lower approximation is evaluated for conditional attributes, while the upper approximation is evaluated for decision attributes. For the upper and lower approximation, a rule is generated automatically. Using the data's boundary region, the combined value from the upper and lower approximations generates the rule. Three rules are consolidated, and the data is clustered. The cluster rule is used to calculate the minimum and maximum value of the conditional attributes. Finally, the RMMC algorithm is executed, and the min-max disparity is calculated to evaluate the result. The result is then evaluated using the test data. To assess the accuracy of the proposed model, K-fold cross-validation is used. The RSMM algorithm is explained below,

Algorithm 2: Rough Set Min-Max classifier

Input: $(U, A \cup D), \theta$

Output: Predicted Result (PR)

Step 1: To evaluate the neighbourhood relationship among the attributes.

$$f(xi, xj) = \sqrt{\sum_{k=1}^n (Xi - Xj)^2}$$

Form the above notation f defined as a metric function of universe U , and (U, f) is the neighborhood relation... $\theta(xi) = \{f(xi, xj) \leq \theta, x \in U\}$

Step 2: Construct the equivalence relation for the decision attribute.

$$IND(Dx) = \{ \forall d \in Dx, d(x) = d(y) \}$$

Step 3: Apply “ $\wedge$ ” (“and”) operator of the neighborhood granules.

$$\theta(Ai \wedge Aj) = \{f(Ai, Aj) \leq \theta, Ai \in U, Aj \in U\}$$

Step 4: Construct the neighbourhood rough set lower approximation ($\underline{ND}_x$), upper approximation ($\underline{ND}_x$) and boundary region ($BNR(D)$) space for decision attributes *for all* D_x

Step 5: Generate certain rules using Neighborhood rough set based on lower approximation ($\underline{NR}$), upper Approximation ($\underline{NR}$) and Boundary region (BR).

Step 6: Consolidate the rules by merging lower, upper, and boundary rules,

$$R = \underline{NR} + \underline{NR} + BR$$

Step 7: For Every Decision, Attribute applies the Min-Max Rule Reduction algorithm.

$$RR = \{ \forall Di \}$$

Step 8: Predict Results for test data using below Min Max Mean Disparity Classifier

$$dpv = \frac{\sum_{i=0}^n \left| T_i - \left(\frac{r_{i \min} + r_{i \max}}{2} \right) \right|}{N_a(R)}$$

$$Ti = i^{th} \text{ attribute in test data}$$

$$r_{i \min} - \text{Minimum value of } i^{th} \text{ attribute in } RR$$

$$r_{i \max} - \text{Maximum value of } i^{th} \text{ attribute in } RR$$

$$N_a(R) = \text{No of attributes in } RR$$

Using a ruleset, the attributes, values, and similarity index group the data into a single group. The Min-max value of the features is calculated and applied to the Min-max disparity in a Rule cluster. Before transforming to the Classification part, the dataset is divided into training and testing data in the ratio of 70:30. Finally, the outcome of the test data is predicted. The RMMC algorithm's accuracy is assessed using k-Fold cross-validation. There are numerous advantages to using the ruleset method, which can process any type of data to predict the output. To reduce the rules, the target reduction technique is used. As a result, the ruleset is small in size, reducing the complexity of the mathematical calculation. This greatly contributes to improving the accuracy of the proposed model.

IV. RESULT AND DISCUSSION

To evaluate the simulation result, this research work used Windows 10 OS with an i5 processor and 4 GB of RAM. Python is the language used to implement IHHO-RMMC for predicting the disease class.

IV.1 DATASET

Three different real-world medical datasets are used to evaluate the performance of the HHO-RMMC algorithm. Epileptic Seizure Recognition Data Set [18], Liver cancer dataset [19], and Breast Cancer dataset [20] are utilized. To predict an accurate result, the proposed classification algorithm is trained and tested k times.

IV.2 PERFORMANCE ANALYSIS

The aforementioned IHHO-RMMC model is implemented in Python, and its performance is evaluated. The training and testing datasets are then divided into a 70:30 ratio. To extract the best features, the training data is fetched to the IHHO optimization algorithm. Following that, the selected features are fed into the RMMC algorithm, and a rough set rule is generated automatically based on the dataset values. It establishes the upper and lower bounds of approximation. The rough set rule is generated, and then the database is divided into test and training data. The ruleset is reduced based on the consolidation rules. Figure 3 depicts the IHHO-RMMC model's work performance. The training data is processed and used for data testing. The accuracy of the IHHO-RMMC is 98.12% higher than that of the existing RMMC model for the breast cancer dataset. The results show that the IHHO-RMMC accuracy outperforms the existing model. According to Fig. 3, the proposed model's performance is calculated for three different datasets: breast cancer, liver, and Epileptic Seizure Recognition Data Set. The proposed model's performance is evaluated using metrics. The IHHO-RMMC and existing algorithm data are calculated using True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values.

Table 1: Performance Analysis of the proposed classification algorithm and other comparative algorithms.

Medical Dataset	Classification algorithm	Precision	Recall	F - measure	Fowlkes Mallows index	Kulczynski index	Rand index
Epileptic Seizure Recognition Data Set [18]	IHHO-RMMC	0.998	0.98	0.98	0.988	0.982	0.99
	RMMC	0.996	0.96	0.97	0.978	0.978	0.98
Liver Dataset [19]	IHHO-RMMC	0.998	0.98	0.98	0.98	0.98	0.99
	RMMC	0.996	0.96	0.97	0.97	0.97	0.96
Breast Cancer Dataset [20]	IHHO-RMMC	0.90	0.98	0.92	0.943	0.94	0.79
	RMMC	0.87	0.97	0.91	0.923	0.92	0.77

Source: Authors, (2026).

IV.3 PERFORMANCE METRICS

1. Precision

To calculate the precision, the TP is divided by the summation of TP and FP value which is represented in equation (15).

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (15)$$

2. ReCa

To calculate the recall, the TP is divided by the summation of TP and FN value which is represented in equation (16).

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (16)$$

3.F-Measure

F-Measure is used to calculate the proposed model accuracy which is represented in equation (17).

$$F - \text{Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (17)$$

4. Fowlkes Mallows index

Fowlkes Mallows index is the square root of the precision and Recall, which is represented in equation (18)

$$\text{Fowlkes Mallows index} = \sqrt{\text{precision} * \text{Recall}} \quad (18)$$

5. Kulczynski index

Kulczynski index is evaluated via summation of precision and recall, which is represented in equation (19).

$$\text{Kulczynski index} = (1/2)(\text{Precision} + \text{Recall}) \quad (19)$$

6. Rand Index

Rand Index is calculated by the summation of TP and FN and divided by the summation of TP, FN, FP, and TN, which is represented in equation (20).

$$\text{Rand Index} = (\text{TP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (20)$$

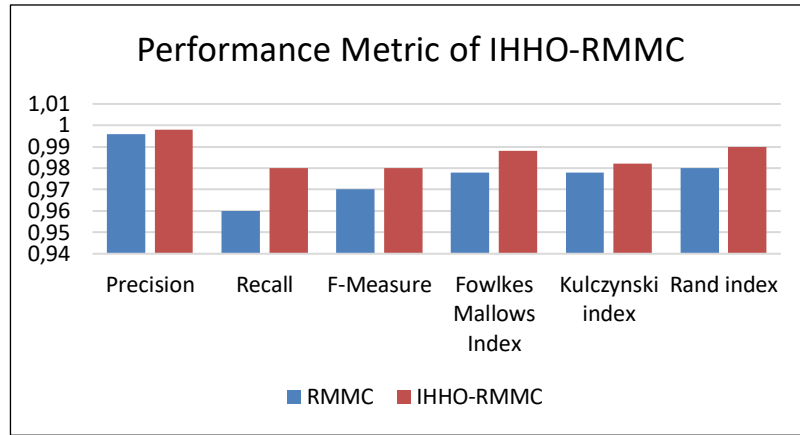


Figure 2: The performance metric of IHHO-RMMC.

Source: Authors, (2026).

Table 2: Performance analysis of the proposed method.

Medical Dataset	Classification algorithm	Accuracy (%)
Epileptic Seizure Recognition Data Set	IHHO-RMMC	99.62
	RMMC	99.42
Breast Cancer Dataset	IHHO-RMMC	98.12
	RMMC	95.22
Liver Dataset	IHHO-RMMC	96.22
	RMMC	94.60

Source: Authors, (2026).

Precision, recall, F-measure, Fowlkes Mallows index, Kulczynski index, and Rand index are calculated for the IHHO-RMMC and RMMC algorithms. The Rand index calculates the clustering value of two similar data sets. The RMMC is 0.98, indicating that the model is of high quality. Unlike the RMMC model, the IHHO-RMMC can easily cluster similar data within the index. The IHHO-RMMC algorithm's performance is measured and compared to that of the RMMC, which is tabulated in table 2 and depicted in Fig 3.

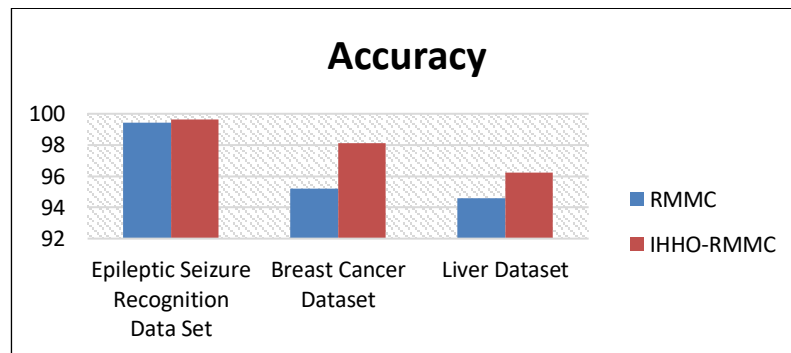


Figure 3: Accuracy of Proposed System.

Source: Authors, (2026).

The proposed method's obtained result is validated based on prediction accuracy. Accuracy is important in medical dataset decision-making and data mining. In the Liver dataset, the IHHO-RMMC algorithm achieves 96.22% more accuracy than the RMMC algorithm (94.60%). Similarly, the IHHO-RMMC achieves 98.12% accuracy in the breast cancer dataset versus the RMMC (95.22%). We tested the IHHO-RMMC algorithm on the Epileptic Seizure Recognition Data Set and found it to be 99.62% accurate, which is depicted in table 2 and fig 3. Following the evaluation of the proposed model, the RMMC model outperforms the NRSC model in terms of accuracy.

V. CONCLUSION

In this paper, we examine disease prediction accuracy using the novel IHHO-RMMC algorithm and enhance feature selection. A fitness function that combines classification accuracy and attributes set size can be used to find feature reductions using IHHO's powerful search capabilities. The IHHO-RMMC algorithm has been modified as a feature selection algorithm for disease prediction. The proposed method uses Improved Lévy Flight Distributions and chaotic maps to enhance the original HHO algorithm to avoid repetitive search by hawks. As part of our future research, we plan to measure the performance of the proposed model using a real IoT dataset.

Future Work: As a part of our future studies, we would like to implement and evaluate the performance of the proposed model in a real IoT dataset. We also desire to explore adaptive parameter updating techniques for dynamically optimizing IHHO-RMMC algorithm. Implementing deep learning techniques in conjunction with feature selection for enhancing disease prediction accuracy even

further will also be included in our extended studies. Besides, deploying the model in a real-time medical monitoring system and studying its computational efficiency in big medical datasets will form a part of our extended studies.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Methodology: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Investigation: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Discussion of results: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Writing – Original Draft: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Writing – Review and Editing: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Resources: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Supervision: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

Approval of the final text: Devi S, V.S. Arulmurugan, V. Sujatha, K.E. Narayana, Saranya K S, M. Balamurugan, J. Parthasarathy.

VII. REFERENCES

- [1] Ge R, Zhou M, Luo Y, Meng Q, Mai G, Ma D, Wang G, Zhou F. McTwo: a two-step feature selection algorithm based on maximal information coefficient. BMC Bioinformatics. 2016;17:142. <https://doi.org/10.1186/s12859-016-0990-0>
- [2] Jun H, Yan-Hong D, Jin-Shu W, Peng Y. Research on the method of warfare simulation data analysis based on BP neural network and RoughSet. Fourth International Conference on Intelligent Systems Design and Engineering Applications. 2013;531-536. <https://doi.org/10.1109/ISDEA.2013.526>
- [3] Nabwey HA. An intelligent mining model for medical diagnosis of heart disease based on rough set data analysis. Int J Eng Res Technol. 2020;13(2):355-363.
- [4] Al-shami TM. Improvement of the approximations and accuracy measure of a rough set using somewhere dense sets. Soft Comput. 2021;25:14449–14460. <https://doi.org/10.1007/s00500-021-06358-0>
- [5] Mafarja MM, Mirjalili S. Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. Soft Comput. 2019;23:6249–6265. <https://doi.org/10.1007/s00500-018-3282-y>
- [6] Tawhid MA, Ibrahim AM. Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm. Int J Mach Learn Cybern. 2020;11:573–602. <https://doi.org/10.1007/s13042-019-00996-5>
- [7] Tawhid MA, Ibrahim AM. Hybrid binary particle swarm optimization and flower pollination algorithm based on rough set approach for feature selection problem. Nature-Inspired Computation in Data Mining and Machine Learning. Springer; 2020. https://doi.org/10.1007/978-3-030-28553-1_12
- [8] Hamed A, Nassar H. Efficient feature selection for inconsistent heterogeneous information systems based on a grey wolf optimizer and rough set theory. Soft Comput. 2021;25:15115–15130. <https://doi.org/10.1007/s00500-021-06375-z>
- [9] Jothi G, Inbarani HH, Azar AT, et al. Rough set theory with Jaya optimization for acute lymphoblastic leukemia classification. Neural Comput Appl. 2019;31:5175–5194. <https://doi.org/10.1007/s00521-018-3359-7>
- [10] Sathiyabhama B, Kumar SU, Jayanthi J, et al. A novel feature selection framework based on grey wolf optimizer for mammogram image analysis. Neural Comput Appl. 2021;33:14583–14602. <https://doi.org/10.1007/s00521-021-06099-z>
- [11] Bhattacharya A, Goswami RT, Mukherjee K. A feature selection technique based on rough set and improvised PSO algorithm (PSORS-FS) for permission-based detection of Android malwares. Int J Mach Learn Cybern. 2019;10:1893–1907.
- [12] Rajesh T, Malar RSM, Geetha MR. Brain tumor detection using optimization classification based on rough set theory. Cluster Comput. 2019;22(Suppl 6):13853–13859. <https://doi.org/10.1007/s10586-018-2111-5>
- [13] Cervante L, Xue B, Shang L, Zhang M. A dimension reduction approach to classification based on particle swarm optimization and rough set theory. AI 2012: Advances in Artificial Intelligence. 2012;7691. https://doi.org/10.1007/978-3-642-35101-3_27
- [14] Bhattacharya A, Goswami RT, Mukherjee K. A feature selection technique based on rough set and improvised PSO algorithm (PSORS-FS) for permission-based detection of Android malwares. Int J Mach Learn Cybern. 2018. <https://doi.org/10.1007/s13042-018-0838-1>
- [15] Zhang Y, Liu R, Wang X, Chen H, Li C. Boosted binary Harris hawks optimizer and feature selection. Eng Comput. 2020. <https://doi.org/10.1007/s00366-020-01028-5>
- [16] Too J, Liang G, Chen H. Memory-based Harris hawk optimization with learning agents: a feature selection approach. Eng Comput. 2021. <https://doi.org/10.1007/s00366-021-01479-4>
- [17] Udhaya Kumara S, Inbarani HH. A novel neighborhood rough set based classification approach for medical diagnosis. Elsevier. 2015.
- [18] UCI Machine Learning Repository. Epileptic Seizure Recognition Dataset. Available from: <https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition>
- [19] Kaggle. Breast Cancer Wisconsin Dataset. Available from: <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
- [20] Kaggle. Liver Tumor Segmentation Dataset. Available from: <https://www.kaggle.com/datasets/andrewmvd/liver-tumor-segmentation>