FMDB *Transactions on Sustainable Health Science Letters*

Random Competition Wild Horse Optimizer and Improved Deep Belief Network Framework for Chronic Disease Detection

R. Arulmathi[1,*], M. Sakthivanitha[2]

[1]Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.
[2]Department of Computer Applications, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.
devmathi@gmail.com[1], sakthivanithamsc@gmail.com[2]

Abstract: The prevalence of Chronic Disease (CD) has increased worldwide, affecting all socioeconomic groups and extending to all regions. The World Health Organisation (WHO) states that early prevention is crucial for some CD, including diabetes mellitus, stroke, cancer, heart disease, kidney failure, and hypertension. Data quality and predictor selection, including machine learning, are key to CD prediction. Data quality is most affected by missing values, normalisation, feature selection, and imbalance. The purpose of CD prediction is to improve performance and address concerns in medical datasets. This research presents a new Deep Learning (DL) CD prediction framework. The Chronic Disease Progression Tracker Dataset is its source. In K-Nearest Neighbor (KNN)-based missing data imputation, missing values are replaced with the values of their k-nearest neighbors. Z-score normalization scales each dataset value to 0 and 1. RCWHO selects the optimal dataset features. Notable improvements in RCWHO include the Competition Waterhole Mechanism (CWM) and the Random Running Approach (RRA). CWM is suggested to improve the CD dataset's best-feature-selection exploitation behaviour. RRA balances exploration with exploitation. ISMOTE generates synthetic samples from the minority class to correct class imbalance in datasets. IDBN classifiers are advanced generative models with positive-etheric distributions for CD classification. Precision, recall, F1-score, accuracy, and error are used to compare classification algorithms in the Matrix Laboratory R2021a (MATLABR2021a) simulator.

Keywords: Chronic Disease Prediction; Missing Data Imputation; Data Normalisation; Feature Selection; Classification and Optimisation; Machine Learning; Deep Learning.

Received on: 27/04/2025, **Revised on:** 30/06/2025, **Accepted on:** 21/09/2025, **Published on:** 07/05/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2026.000636

Cite as: R. Arulmathi and M. Sakthivanitha, "Random Competition Wild Horse Optimizer and Improved Deep Belief Network Framework for Chronic Disease Detection," *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 2, pp. 102–118, 2026.

1. Introduction

In the world, the prevalence of chronic diseases has increased, impacting all socioeconomic levels and spreading across all geographical areas. Arthritis, asthma, back pain, cancer, cardiovascular disease, chronic obstructive pulmonary disease, diabetes, chronic kidney disease, mental health disorders, and osteoporosis are the 10 primary categories of chronic conditions. Chronic illnesses are a major global healthcare issue, leading to rising death rates and consuming over 70% of patients' income

*Corresponding author.

on therapies. To reduce risk factors, healthcare data should include demographics, medical history, and living situation, as well as environmental conditions, to understand disease causes [1]; [2]. The integration of information technology (IT) has driven significant changes in the healthcare sector over the past few years. Since smartphones have made people's lives easier, the goal of integrating IT into healthcare is to improve people's lives and reduce costs [3].

Making healthcare more intelligent, for example, by creating smart hospital buildings and smart ambulances—could make this possible and benefit both patients and physicians in several ways [4]. A yearly study of patients with CD in a particular area identified many patients and very little gender variation. Combining structured and unstructured data in healthcare yields more accurate findings, as unstructured data includes medical records, patient symptoms, and complaints [5]; [6]. CDs are caused by risk behaviours such as high blood pressure, tobacco use, elevated cholesterol, poor diet, inactivity, and excessive alcohol consumption. Rare diseases are difficult to diagnose [31]; [36]. Because the processes being performed (feature selection (FS), dimensionality reduction, and classification) depend on the chosen database, the database should be appropriate for these processes.

The process of selecting the more significant and pertinent features from the datasets is known as FS. FS is important and has grown significantly because it can improve learning accuracy, reduce learning time, and simplify learning outcomes [33]. To increase classification accuracy, the FS method identifies the most important and relevant features and reduces them without significantly compromising classification performance [34]. Removing redundant and unnecessary attributes lowers data dimensionality and improves classification efficiency [35]. Preprocessing methods, such as feature selection and dimensionality reduction in data mining, reduce data size, improve visualisation, and reduce training time for classification systems diagnosing CD.

The feature selection methods used in data mining can be categorised into three types: Filter, wrapper, and embedding techniques [7]; [8]. To select the best feature subset, filter models consider the data's statistical properties [9]; [10]. Traditional FS approaches, such as Information Gain, Chi-square, and ReliefF, are also known as filter models. Two other categories of filter models are univariate feature filters, also called feature ranking algorithms. Each feature is assigned a weight in feature ranking algorithms based on its relevance to the target notion. On the other hand, multivariate filters are techniques for subset search. While the wrapper model uses optimisation methods such as meta-heuristics (MH), which directly communicate with the features and classifier, the filter model cannot establish a direct connection with the classifier [11]; [12].

Furthermore, like the wrapper model, embedded models incur lower computational cost when interacting with the classifier [37]. To improve classification accuracy, FS approaches eliminate redundant and irrelevant information from the original database [38]. The machine learning problem of feature selection is challenging because assessing the model's performance at every subset is difficult, especially when n is large. Techniques such as exhaustive, greedy, and random searches have been used to identify optimal subsets in feature selection problems. Most of the approaches have the highest computing cost, high complexity, and premature convergence [39]. As a result, MH algorithms receive significant attention for handling such situations.

They are the most effective and efficient methods, and they can identify the optimal subset of features while preserving the model's accuracy. Numerous MH algorithms have been developed over the past three decades to address various optimisation problems [13]. Machine learning is a subset of artificial intelligence techniques that use data to simulate intelligent behavior without human involvement [14]. AI in prescriptions will automate work, freeing medical professionals' time for other responsibilities. ML can be categorised into unsupervised and supervised models [15]. By identifying intricate models and extracting medical knowledge, ML introduces practitioners and experts to new concepts [16].

ML prediction models can be used in clinical practice to identify improved parameters for making decisions regarding the care of specific patients. Since deep learning offers a more effective learning mechanism for classification issues than traditional ML techniques, it is being successfully employed in a variety of fields. However, because medical data is sparse, varied, and unstructured, applying deep learning techniques to it remains difficult. An artificial neural network called a Deep Belief Network (DBN) is utilised for unsupervised learning tasks such as generative modelling, dimensionality reduction, and feature learning. This system, which learns to represent data hierarchically, comprises multiple layers of hidden units. DBNs are used across a range of domains, including collaborative filtering, natural language processing, and image identification [17]. In this paper, an Improved Deep Belief Network (IDBN) approach is introduced for diagnosing chronic illnesses.

The Random Competition Wild Horse Optimizer (RCWHO) is presented, which uses two distinct operators, namely the CWM and the RRA, to determine their optimal features. To find the best global solution, RRA is recommended to balance exploration and exploitation. To improve exploitation behaviour, the CWM is added to stallion position computations. IDBN is a probabilistic graphical model for CD classification that consists of several layers of hidden units. Usually, IDBN is trained layer by layer using Restricted Boltzmann Machines (RBMs), which help initialise the network's weights and address the

vanishing gradient problem in deep networks. IDBN aims to enhance diagnostic accuracy and classification performance. To compare classification algorithms using performance measures (precision, recall, F1-score, accuracy, and error), the MATLAB R2021a (MATLABR2021a) simulator is utilized.

2. Literature Review

Lu and Uddin [18] proposed a Graph Neural Network (GNN) for CD prediction. To extract the latent association among patients, a Weighted Patient Network (WPN) is created by projecting a patient-disease bipartite graph. GNNs are used to build predictive models that provide reliable patient representations for CD prediction using WPN features. The GNN framework will improve CD prediction accuracy. Additionally, the pattern for the two cohorts is displayed in the visualisation of the final hidden layers of GNN-based models, highlighting the discriminative strength of the proposed model. The Commonwealth Bank Health Society (CBHS), an Australian health fund corporation, provided the administrative claim data.

Zhang et al. [19] introduced a Convolutional Neural Network (CNN) for CD prediction. GroupNet transforms CD prediction into a multi-label classification problem. The Binary Relevance (BR) and Label Powerset (LP) approaches are used to transform several CD labels. GroupNet incorporates correlations among several diseases via a correlated loss. A patient may be diagnosed with all three chronic diseases, a different combination of the three, or no evidence of any of the three, according to the BR approach. The Scikit-learn package, the Tensorflow platform, and the Waikato Environment for Knowledge Analysis (WEKA) software serve as the foundation for all investigations. The physical examination datasets were collected from a nearby medical facility for the studies. GroupNet performs better than the others according to the evaluation metrics.

Jain and Singh [20] created the Support Vector Machine (SVM), Relief, and Principal Component Analysis (PCA). The suggested strategy is a hybrid approach combining the Relief method and PCA feature selection with an optimised SVM classifier. Lung Cancer (Harvard), Leukemia Cancer, Lung Cancer (Michigan), Ovarian Cancer, Prostate Cancer, Colon Dataset, Heart Disease, Chronic Kidney Disease, and Pima Indian Diabetes Dataset are nine well-known disease datasets used for medical diagnosis to assess the model's effectiveness. With notable reductions in dimensionality and computational complexity, the resulting system exhibits improved classification accuracy and can adapt to any dataset type. Furthermore, when compared to adaptive SVM without feature selection and conventional SVM, the system shows greater accuracy. The suggested model is evaluated using several metrics, including F-Measure, Accuracy, Recall, Precision, and Specificity.

Gabriel and Anbarasi [21] presented an Optuna hyperparameter tuning of eXtremely Gradient Boost (BSOXGB) based on BorutaShap feature selection. The proposed approach, including Data Preprocessing (DP), feature selection strategies, and Hyperparameter tuning (HP), is introduced for the diagnosis of coronary artery disease (CAD) to achieve optimal performance and accuracy on these datasets. By removing noise, managing missing numbers, and handling outliers, DP is also essential to achieving accurate results. Following feature selection, BSOXGB was created for hyperparameter tuning and optimisation. In the Z-Alizadeh Sani dataset, the suggested model outperforms existing classifiers such as Random Forest (RF), AdaBoost (AB), CatBoost (CB), and ExtraTrees (ET), and achieves the highest accuracy. BSOXGB achieves the highest classification accuracy with only 9 of 56 relevant features, suggesting it could be a viable method for automatically identifying and diagnosing CAD in the real world.

The dataset was obtained from the University of California, Irvine (UCI) machine learning repository. Several metrics, including accuracy, precision, recall, F1-score, specificity, log loss, and the Area Under the Curve-Receiver Operating Characteristic (AUC-ROC), are used to assess the classifier's performance. Mishra et al. [22] developed the following methods, such as Correlation Feature Selection (CFS), Information Gain (IG), Chi-Square (CS), Best First Search (BFS), Linear Forward Selection (LFS), Greedy Stepwise Search (GSS), and Decision Tree (DT) for the prediction of CD. Dataset collection, attribute selection, and classification are the suggested work steps. To begin, datasets were gathered from the UCI repository.

The second is a comparative study of attribute selection that examines how filter and wrapper selection techniques affect classification performance. The WEKA tool is then utilised to develop the DT method, which is employed as a classifier in this investigation. Datasets on diabetes, breast cancer, and heart disease have all undergone attribute significance analysis. Additionally, an enhanced DT classifier was used to assess the hybrid approach. Its performance was documented and validated across 14 distinct CD datasets using sensitivity, specificity, F1 score, and accuracy. Lambert and Perumal [23] introduced a Deep Neural Network (DNN) and Firefly Optimisation (FFO) for the classification of chronic renal disease. A novel classification model for the diagnosis of CKD that uses optimal feature selection based on metaheuristic algorithms. Initially, during preprocessing, missing values were imputed [30].

The optimal feature selection is then determined using a metaheuristic algorithm called the Oppositional-based FFO method, which improves disease classification accuracy. The FFO rate of convergence is increased with the use of the oppositional-based learning principle. DNN is finally introduced to diagnose the presence of CKD. According to the findings, the suggested

algorithm outperformed the current classifier models in terms of accuracy, sensitivity, and specificity. Al-Jamimi [24] used Extreme Gradient Boosting (XGBoost) and Recursive Feature Elimination with Support Vector Machine (RFE-SVM) to predict CD. Heart, breast cancer, diabetes, and kidney cancer are the first datasets to undergo preprocessing. To reduce data complexity, the methodology provided identifies and eliminates unimportant features. Next, RFE-SVM is used to choose features. The powerful XGBoost classifier, renowned for its ability to capture intricate correlations in the data, is then fed the modified dataset. The selected ensemble learning algorithm is very effective at generating complex patterns, which are essential for predicting chronic diseases.

By adjusting hyperparameters using Bayesian optimization, a crucial optimization phase is introduced to improve model performance. By carefully navigating the hyperparameter space, this optimizes search efficiency and fine-tunes the model to achieve the best possible predictive accuracy. Accuracy, precision, recall, F1-score, AUC-ROC, and Matthew's Correlation Coefficient (MCC) are the evaluation measures. Singh et al. [25] presented a Particle Swarm Optimisation with Random Forest (PSORF) for the automatic detection of CDs. Firstly, the model was trained using data preprocessing, SMOTE, and Expectation Maximization (EM) imputation methods. Secondly, PSORF is the RF classifier for multiple CDs, and PSO is used to find the smallest optimal feature set, thereby optimizing the RF classifier's performance.

Thirdly, CD classification, five distinct CD datasets have been used as evaluation benchmarks. The findings demonstrated that across all five datasets, the suggested technique had the highest mean rank for accuracy, F-measure, and ROC. Yuan et al. [26] created a Network-Limited Polynomial Neural Network (NLPNN) for early diagnostic CDs. By effectively capturing high-level features hidden in CD datasets, the NLPNN algorithm augments the data in its feature space and prevents overfitting. Then, the attention-empowered NLPNN algorithm, which also augments the data's sample space, is applied to improve diagnostic accuracy. Nine public and two real CD datasets, partially with class imbalance, are used to simulate the suggested framework.

The results of numerous experiments show that the proposed diagnostic algorithms outperform the most advanced algorithms, achieving higher recall, F1-score, G-mean, and accuracy. Daid et al. [27] developed a Multilayer Perceptron (MLP) to diagnose the CD. A multilayer convolution deep learning methodology, a type of deep learning model that treats the input medical data as a vector representation, is used to forecast the chronic diseases of various individuals. Furthermore, MLP is a feed-forward NN that primarily consists of three processing layers for diagnosing CD. Based on the process for determining whether a patient may have a high CD rate due to chronic illness, the suggested system achieved effective disease prediction.

The deep learning model has a low classification error rate and higher precision and accuracy in predicting several diseases. The CD dataset, which includes symptoms of depression, exhaustion, headaches, and other body aches, is used in the proposed work to evaluate a useful methodology for managing and forecasting chronic diseases using partially experimental data. Mishra et al. [28] developed a Memory-based Metaheuristic Attribute Selection (MMAS) using K-means clustering (K-means) and Naive Bayes (NB) for CD risk assessment. MMAS is introduced, which optimizes data on chronic illnesses by performing a local neighborhood search.

Unsupervised K-means clustering is used to filter it further and eliminate outliers. The NB classifier uses the resulting data to identify the presence of chronic disease risks. Thus, heuristic-based attribute selection, combined with clustering and classification, can greatly enhance the diagnosis of chronic diseases. The datasets used include hepatitis, diabetes, breast cancer, and heart disease. Compared with alternative state-of-the-art methods, the proposed model achieved higher precision, recall, and F-score. A decision support system will help medical professionals to make CD diagnoses more proficiently and efficiently.

3. Proposed Methodology

In this study, data is collected from the Chronic Disease Progression Tracker Dataset. K-Nearest Neighbour (KNN) based missing data imputation is introduced to replace or impute values in a dataset. Z-score normalization, which scales each value in a dataset to a mean of 0 and a standard deviation of 1, is used to standardize the dataset. The RCWHO, which combines CWM and the RRA, is presented for choosing the optimal features from the dataset. The CWM is proposed to enhance exploitation behaviour in the best-feature-selection procedure for the CD dataset. At the same time, the RRA is used to discover an equilibrium between exploration and exploitation.

ISMOTE generates synthetic samples for the minority class to reduce class imbalance. An advanced generative model with a positive-etheric distribution for CD classification is the Improved Deep Belief Network (IDBN). Precision, recall, f1-score, accuracy, and error are compared to classification techniques using the MATLAB R2021a simulator. The flow process of the proposed model is displayed in Figure 1.

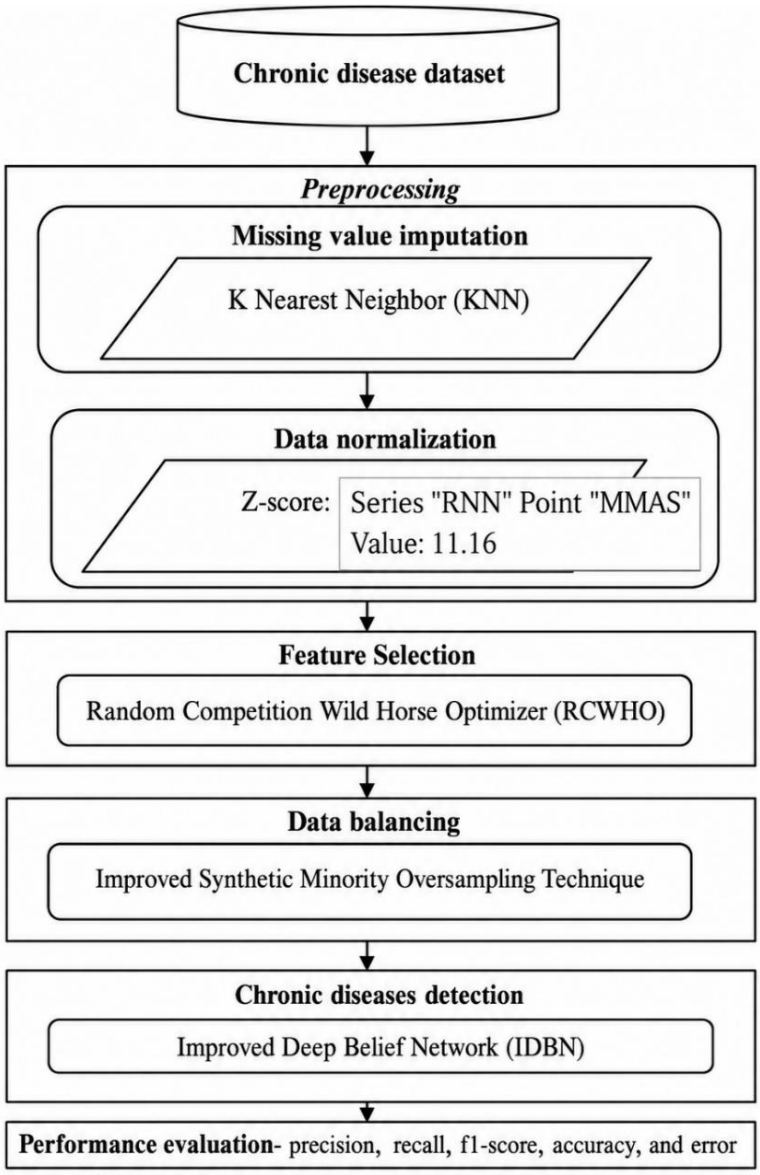

Figure 1: Overall flow of proposed model

3.1. K-Nearest Neighbor (KNN) Based Missing Data Imputation

KNN is a method that uses similarity between samples to fill in missing values in a dataset. Nearest-neighbour distances are computed using the Euclidean distance to impute missing values in the dataset. Finding the "K" closest neighbours of a sample is used to replace missing values by imputing them with data from those neighbours. Taking the mean or median of the nearby numbers is a common method. To accomplish this, it assigns higher weights to non-missing values and ignores missing values when calculating the Euclidean distance. Equation (1) can be used to determine the Euclidean distance [29]:

$$D_{xy} = \sqrt{weight * squared\ distance\ from\ present\ coordinates} \quad (1)$$

Where weight is the amount of all coordinates currently.

3.2. Z-score Normalisation-Based Data Normalization

Normalising each value in a dataset is called Z-score normalisation. The feature values are transformed using the considered feature's meaning and standard deviation. More precisely, the following equation (2) is used to convert values for the considered feature into new normalised values [31]:

$$\mathbf{v}' = \frac{\mathbf{v}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} \quad (2)$$

Z-score normalisation technique, σ is the standard deviation, the mean value (μ) of a feature is negative, above is positive, and zero for values equal to the mean.

3.3. Random Competition Wild Horse Optimiser (RCWHO) based Feature selection

The WHO algorithm is inspired by the social behaviours of horses, including mating, dominance, grazing, and leadership hierarchy. For the best feature selection from the CD dataset, horses can generally be categorized into two classes (territorial and non-territorial) based on their social organization. They dwell in groups of various ages, including mares, stallions (S), and progeny. For the best feature selection from the CD dataset, mares and stallions coexist while engaging in mutual interactions during grazing. After growing up, foals leave their groups to form their own families, selecting the best features from the CD dataset. For optimal feature selection, this behaviour prevents siblings and stallions from mating. The WHO algorithm has five steps, which are explained below [31].

Population Initialization: The number of leaders is G, and the number of non-leaders (mares and foals (F)) is N–G, where N is the total number of individuals and G is the number of groups. PS, which is G/N, is the percentage of stallions.

Grazing Behaviour: For the best feature selection from the CD dataset, a foal spends the majority of its life grazing close to its group. Equation (3) is utilised to allow other people to move:

$$X_{G,j}^{i} = 2Z\cos(2\pi rZ) \times \left(\mathcal{S}_{G,j} - X_{G,j}^{i}\right) + \mathcal{S}_{G,j} \quad (3)$$

Where $X_{G,j}^{i}$and $\mathcal{S}_{G,j}$is described as the ith group member and $\mathcal{S}$ in the jth group, $r \in (-2,2)$ is a random value, and Z is computed using Equation (4):

$$P = \overrightarrow{r_1} < TDR, IDX = (P == 0), Z = r_2 \Theta IDX + \overrightarrow{r_3}\Theta(\sim IDX) \quad (4)$$

$P \in (0,1)$ is a random vector, $\overrightarrow{r_1}, r_2, \overrightarrow{r_3} \in (0,1)$ are random vectors. TDR is a linearly past its best metric using equation (5):

$$TDR = 1 - \frac{t}{\mathrm{T}} \quad (5)$$

Where t and T are denoted as the present and maximum iterations.

Horse Mating Behavior: As previously mentioned, removing F from their original groups before they reach maturity and mate is one of the distinctive behaviors that horses exhibit compared to other animals. Equation (6) is utilised to allow for the simulation of horse mating behaviour:

$$X_{G,k}^{p} = Cr\left(X_{G,j}^{q}, X_{G,j}^{z}\right), i \neq j \neq k, q = z = end \quad (6)$$

$$Cr = Mean$$

Where $X_{G,k}^{p}$is horse p location in group k through q, and z locations of horses in groups i and j. P_{Cr} is the probability of crossing is set at a constant in the RCWHO.

Group Leadership: Group leaders (\mathcal{S}) will guide the others to a waterhole or other appropriate location. The dominant group will use the waterhole first for the best feature selection, as \mathcal{S} will also strive for it. This behaviour is simulated using equation (7):

$$\overline{\mathcal{S}}_{G,j} = \begin{cases} 2Zcos(2\pi rZ)\left(WH - \mathcal{S}_{G,j}\right) + WH \; if \; rnd > 0.5 \\ 2Zcos(2\pi rZ)\left(WH - \mathcal{S}_{G,j}\right) - WH \; if \; rnd \leq 0.5 \end{cases} \quad (7)$$

Where $\overline{\mathcal{S}}_{G,j}$ and $\mathcal{S}_{G,j}$are the leaders' current location and the candidate location in the jth group, and WH is the waterhole location.

Exchange and Selection of Leaders: Leaders are randomly generated and elected depending on their fitness. To model the interaction among leaders and extra individuals by equation (8),

$$\mathcal{S}_{G,j} = \begin{cases} X^i_{G,j}, if\ f\left(X^i_{G,j}\right) < f\ (\mathcal{S}_{G,j}) \\ \mathcal{S}_{G,i}, iff\left(X^i_{G,j}\right) \geq f\ (\mathcal{S}_{G,j}) \end{cases} \quad (8)$$

Where the fitness values of stallion and foal are represented by $f(\mathcal{S}_{G,j})$ and $f(X^i_{G,j})$ [32]. WHO is added to improve the model's optimization capacity by introducing two techniques. First, in the wild, horses usually chase and run. The F and $\mathcal{S}$ are subjected to a random running strategy. Finally, a waterhole competition is proposed to enhance the selection of stallions.

Random Running Approach (RRA): Wild horses enjoy running and chasing one another. This serves as inspiration for the random running approach, which is suggested for both F and \mathcal{S}_\ with optimal features. The following is the structure of the position-updating equation (9):

$$\boldsymbol{X^i_{G,j}\ or\ \overline{\mathcal{S}}_{G,j} = l_b + (ub - l_b) \times rand} \quad (9)$$

Where the lower and upper boundaries are denoted by l_b and u_b, respectively, this technique might help search agents escape local optima. To achieve optimal feature selection in the CD dataset, the probability of random running (PRR) is set to 0.1 to balance exploration and exploitation.

CWM: Furthermore, two \mathcal{S} can locate the waterhole at the same time. Therefore, the CWM is suggested as a replacement for the second Equation (7) to enhance the quality of the solution further. The following is the position update equation (10):

$$\overline{\mathcal{S}}_{G,j} = WH - Z \times \left(\mathcal{S}_{G,j} \times Q_1 - \mathcal{S}_{G,j} \times Q_2\right) \quad (10)$$

Where Q_1 and Q_2 is denoted by a random number between -1 and 1. RRA and CWM, the following equation (11) is used to replace equation (7) in CWHO:

$$\overline{\boldsymbol{\mathcal{S}}}_{\boldsymbol{G,j}} = \begin{cases} \boldsymbol{2Zcos(2\pi rZ) \times \left(WH - \mathcal{S}_{G,j}\right) + WH, if\ r_3 > 0.5} \\ \boldsymbol{WH - Z \times \left(\mathcal{S}_{G,i} \times Q_1 - \mathcal{S}_{G,j} \times Q_2\right), if\ r_3 \leq 0.5} \end{cases} \quad (11)$$

F and $\mathcal{S}$ have more options for updating their positions in RCWHO. Search agents can improve stability, exploration, and exploitation using the RRA. In addition, CWHM accelerates convergence and enables the method to deliver high-quality solutions. Algorithm 1 displays the RCWHO pseudo-code.

Algorithm 1: Pseudo-code of RCWHO

Start

1. Input parameters: $P_{Cr} = 0.13$, PS = 0.2, PRR = 0.1.
2. Set population size (N) and the maximum number of iterations (T)
3. Initialise the population of horses at random.
4. Create F groups and select $\mathcal{S}$.
5. While (t ≤T)
6. TDR is determined by Equation (5)
7. Z is determined by Equation (4)
8. For the amount of $\mathcal{S}$
9. For the amount of F
10. If rand > PC
11. F location is updated by Equation (3)
12. Else if rand > PRR
13. F location is updated by Equation (6)
14. Else
15. F location is updated by using Equation (9)
16. End if
17. End for F
18. If rand > PRR
19. If rand > 0.5
20. Candidate location of $\mathcal{S}$ is created by Equation (10)

21. Else if
22. Candidate location of $\mathcal{S}$ is created by Equation (11)
23. End if
24. else if
25. Candidate location of $\mathcal{S}$ is created by Equation (9)
26. end if
27. If the candidate location of the $\mathcal{S}$ is better
28. Swap the location of the $\mathcal{S}$ using the candidate location.
29 End if
30. End for $\mathcal{S}$
31. F and $\mathcal{S}$ are exchanged location by Equation (8)
32. t = t + 1
33. End While
34. Output the best optimal feature by RCWHO

End

3.4. Improved Synthetic Minority Oversampling Technique (ISMOTE) Based Data Balancing

In the SMOTE algorithm, each minority class sample is oversampled equally with other samples. However, it is not always required to assign each sample the same weight. Because they are challenging to classify, the authors believe that samples far from the cluster must be given a high priority [33]; [34]. After feature selection, split the minority samples and apply the Improved-SMOTE algorithm to them. There are N samples and W features in the minority CD dataset. The CD dataset's minority samples are now used to create a Euclidean matrix. The matrix dimension can be expressed as [N, W]. Equation (12), Euclidean distance (ED), is used to determine each sample's distance from the others. A matrix of dimension [N, N] is produced after looping over:

$$ED(m_i, m_j) = \sqrt{\sum_{Z=1}^{N}(m(I,z) - m(j,z))} \quad (12)$$

Equation (12), $i, j \in [0, N]$, ED each one of the diagonal values (where j = i) is 0.

Probability weight p of each sample in the interval (0, 1) to determine its relative significance while building the oversampling matrix. Consequently, the overall distance of the samples from the minority class cluster is indicated by the sum of the matrix. Equation (13), Euclidean Distance Sum (EDS), is used to represent the ED,

$$\boldsymbol{EDS} = [\sum \boldsymbol{ED_{1i}}, \sum \boldsymbol{ED_{2i}}, \sum \boldsymbol{ED_{3i}}, \ldots\ldots \sum \boldsymbol{ED_{Ni}}] \quad (13)$$

Consequently, normalise the matrix as NM by equation (13):

$$\boldsymbol{NM} = \frac{\boldsymbol{EDS}}{\boldsymbol{sum(EDS)}} \quad (14)$$

The amount of oversampling can be specified by this NM. After rounding the results to two decimal places, multiply them by the oversampling percentage (T%):

$$\boldsymbol{Oversampling} = \frac{\boldsymbol{N*T}}{\boldsymbol{100}}[\boldsymbol{NM}] \quad (15)$$

ISMOTE has been used for classification, and it is a useful method for handling unbalanced datasets.

3.5. Improved Deep Belief Network (IDBN) based Classification

The deep probabilistic graphical model, the Deep Belief Network (DBN), comprises several layers with varying numbers of neurons in the CD dataset. Nodes on the same layer are independent of one another and do not share a link. Two neighbouring layers have fully linked neurons; the visible layer, the lowest layer of the neural network, is used to input data features, while the additional layers represent hidden layers. From top to bottom, the layers are connected. In a DBN with m levels, the CD dataset and its features are input into the lowest layer, also known as the visible layer. The probability of opening variables at each layer in the Restricted Boltzmann Machine (RBM) structure is determined by the bidirectional link between the top two layers of neurons. The likelihood of each layer's variables opening depends on CD variables at the top layer and connections

between neurons, forming a sigmoid confidence block in the dataset. The energy function to thermal equilibrium is expressed in equation (16) as follows:

$$\boldsymbol{E(v,h) = -a^T v - b^T v - v^T wh} \tag{16}$$

The bias of visible and hidden units is represented by a and b, whereas the weight between them is represented by v, h, and w. The conditional distribution is sampled repeatedly, and a hidden layer with probability is obtained using the IDBN model:

$$\boldsymbol{p(v,h) = \frac{e^{-E(v,h)}}{\sum_v \sum_h e^{-E(v,h)}}} \tag{17}$$

The likelihood of activation of hidden-layer neurons is determined using equation (18):

$$\boldsymbol{p(h_j|v) = \sigma\left(b_j + \sum_j w_{ij} v_j\right)} \tag{18}$$

Equation (19) provides the likelihood that the hidden-layer neuron will activate the visible-layer neuron:

$$\boldsymbol{p(v_i|h) = \sigma\left(a_i + \sum_j w_{ij} h_j\right)} \tag{19}$$

Where the sigmoid activation function is, the training procedure is illustrated in Figure 2, using σ and the logistic function to ensure the probability density exhibits CD independence.

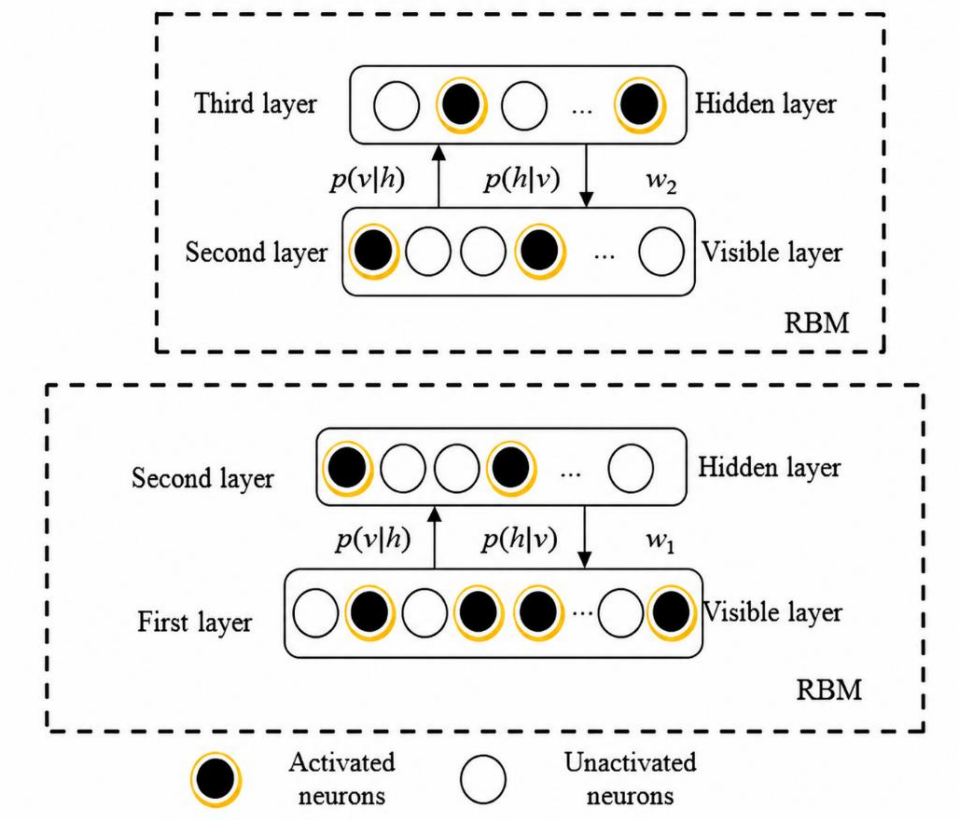

Figure 2: Layer-by-layer pre-training process

Pre-Training Method: Establish the training set, learning rate, network layers, and weights and biases; initialise them to zero; use hidden variables to determine their values and probabilities at each layer; and repeat until reaching the final layer. The final layer is trained using an RBM to determine training weight and bias values, with the implicit variable being a CD dataset. The DBN finely adjusts the weights between layers, combining the upward cognitive and downward generation matrices, to accelerate the model's convergence to a local optimum. Figure 3 illustrates the fine-tuning procedure, focusing on maximising the probability of the upper layer of the upward cognitive weight matrix for sampling.

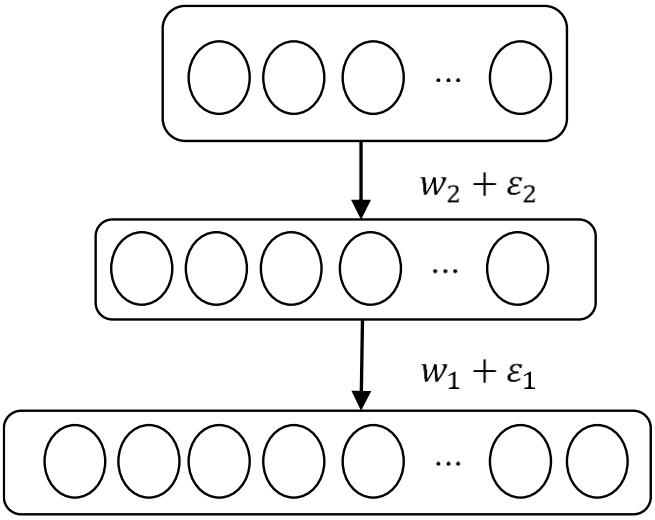

Figure 3: Illustration of the fine-tuning process

The IDBN method is used to select the positive-etheric distribution function at sampling time as described in equation (20):

$$p(x_i = x|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)}{2\sigma^2}\right) \quad (20)$$

Where the current state x_i can be used to derive the parameter μ, which is learned from the RBM network, and equates 1. The explicit layer with conditional probability to the hidden layer in RBM is as follows:

$$\boldsymbol{p(v_i|h) = N\left(a_i + \sum_j w_{ij} h_j, 1\right)} \quad (21)$$

The hidden layer with conditional probability to the explicit layer is as follows:

$$\boldsymbol{p\left(h_j|v\right) = N\left(b_j + \sum_i w_{ji} v_i, 1\right)} \quad (22)$$

The hidden features formed by IDBN are more correct than those produced by a binary RBM.

4. Results and Discussion

In this section, numerous tests were conducted against classification techniques such as GNN, CNNRNN, and IDBN. Additionally, it evaluates the effectiveness and superiority of feature selection techniques, including RCWHO, FFO, BSO, and MMAS. Additionally, the number of persons per FS technique was set to 30, and the maximum number of iterations was set to 500. The performance of IWHO and traditional algorithms was compared. The benchmark CD dataset was used to simulate these techniques. The tests were conducted on an Intel(R) Core (TM) i5-9500 CPU running Windows 10 with 16 GB of RAM. MATLAB R2021a was used for all simulations.

4.1. Dataset

The source of the CD dataset [40], which includes 3000 samples and 26 attributes, focuses on diabetes, Parkinson's disease, and Alzheimer's disease, and documents their evolution over time. For synthetic patients, it contains biometric measurements, pharmaceutical information, lifestyle characteristics, and disease stage data. With several entries per patient and a time-series format, the data is excellent for evaluating risk factors at various stages of chronic illness, modelling disease development, and predicting future biometrics.

4.2. Evaluation Metrics

This section compares classification methods using performance metrics (precision, recall, F1 score, accuracy, and error).

Precision: Precision is the ratio of precisely predicted positive instances to each one of the instances predicted as positive by equation (23):

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ positive}}{\boldsymbol{true\ positive + false\ positive}} \quad (23)$$

Recall: Equation (24) defines recall as the proportion of correctly predicted positive cases among the actual positive instances in the sample:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ positive}}{\boldsymbol{true\ positive + False\ \ Negative}} \quad (24)$$

F1-score: Equation (25) defines the F1-score of the system as the harmonic mean of precision and recall:

$$\mathbf{F1 - Score} = \frac{\mathbf{2 \times precision \times recall}}{\mathbf{precision + recall}} \quad (25)$$

Accuracy: Accuracy describes the portion of predictions that are established properly. Properly, accuracy is described by equation (26):

$$\mathbf{Accuracy} = \frac{\mathbf{True\ positive + True\ Negative}}{\mathbf{Total}} \quad (26)$$

Error: Error is a measure of how frequently a classifier formulates wrong predictions by equation (27):

$$\text{Error Rate} = 1 - \text{accuracy} \quad (27)$$

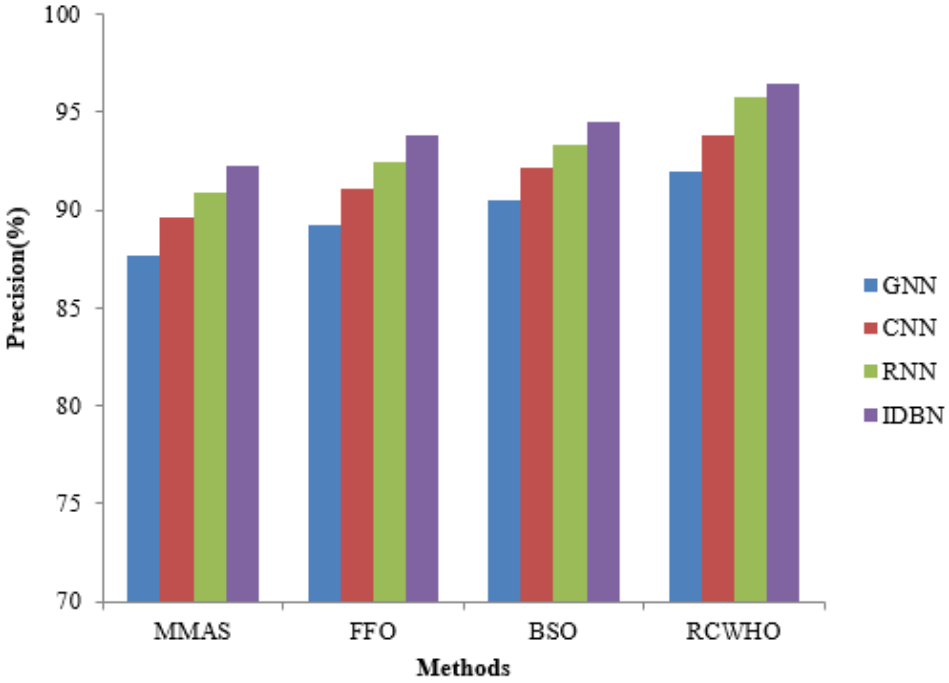

Figure 4: Precision analysis comparison of detection methods

Figure 4 shows the precision analysis of classifiers with feature selection methods, including MMAS, FFO, BSO, and RCWHO. GNN achieves precision of 87.69%, 89.24%, 90.51%, and 91.93%, and CNN achieves precision of 89.65%, 91.07%, 92.18%, and 93.85% for MMAS, FFO, BSO, and RCWHO, respectively. MMAS, FFO, BSO, and RCWHO with respect to RNN achieve precision of 90.88%, 92.46%, 93.37%, and 95.81%, respectively, and IDBN achieves the highest precision of 92.27%, 93.82%, 94.50%, and 96.45% against MMAS, FFO, BSO, and RCWHO, respectively. The proposed IDBN classifier achieves 4.52%, 2.6%, and 0.64% increases in precision over GNN, CNN, and RNN, respectively.

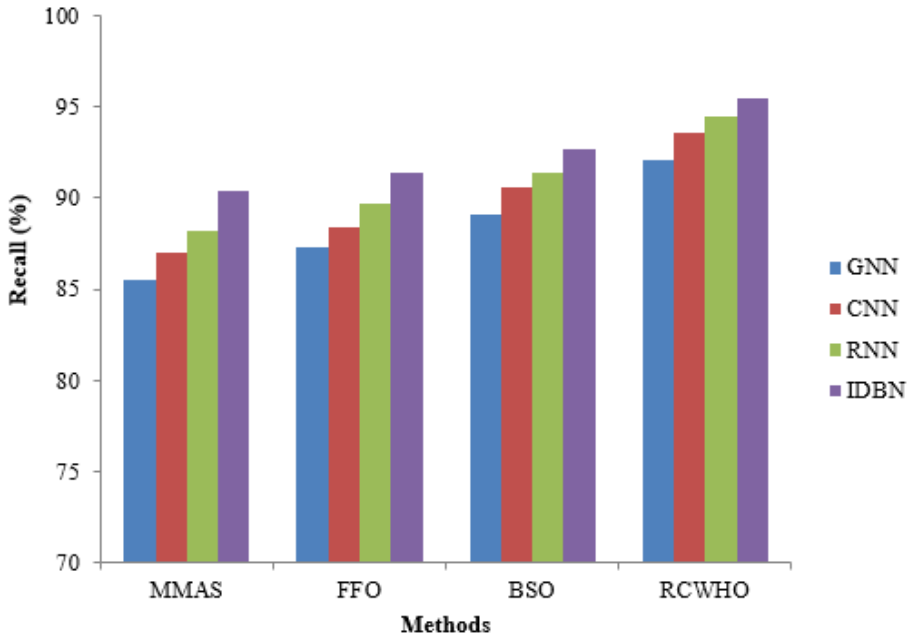

Figure 5: Recall analysis comparison of detection methods

Figure 5 shows the recall analysis of classifiers against MMAS, FFO, BSO, and RCWHO. GNN achieves a recall of 85.47%, 87.35%, 89.08%, and 92.11% against MMAS, FFO, BSO, and RCWHO, respectively, and CNN achieves a recall of 87.03%, 88.41%, 90.64%, and 93.56% against MMAS, FFO, BSO, and RCWHO, respectively.

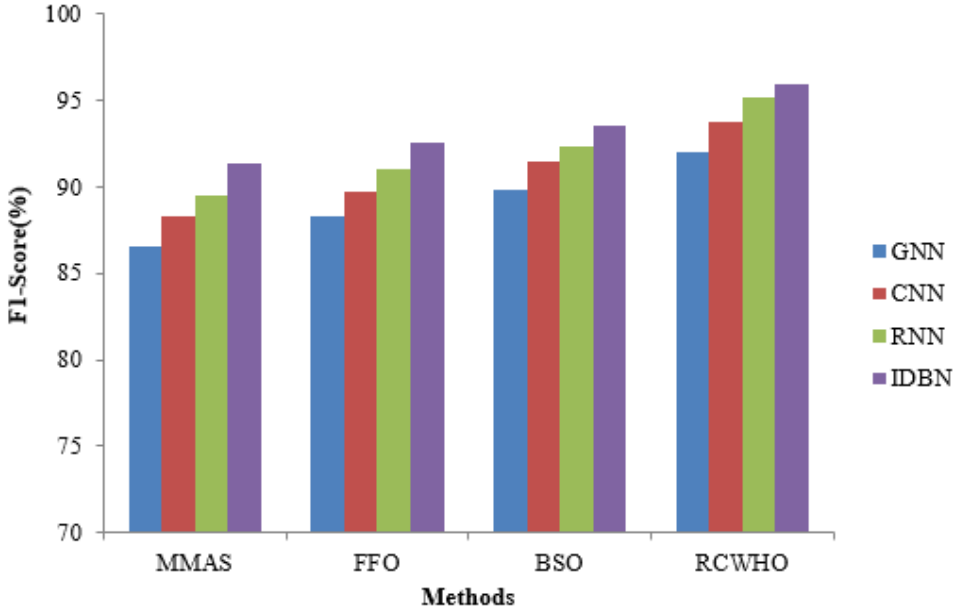

Figure 6: F1-score analysis comparison of detection methods

MMAS, FFO, BSO, and RCWHO with respect to RNN achieve recall of 88.19%, 89.67%, 89.67%, 91.38%, and 94.49%, respectively. IDBN achieves the highest recall of 90.39%, 91.41%, 92.67%, and 95.50% against MMAS, FFO, BSO, and RCWHO, respectively. The IDBN classifier achieves recall increases of 3.39%, 1.94%, and 1.01% over GNN, CNN, and RNN, respectively. F1-score analysis of classifiers against MMAS, FFO, BSO, and RCWHO is illustrated in Figure 6. GNN gives F1-scores of 86.57%, 88.28%, 89.79%, and 92.02%, and CNN gives F1-scores of 88.33%, 89.72%, 91.41%, and 93.70% against MMAS, FFO, BSO, and RCWHO. MMAS, FFO, BSO, and RCWHO, with respect to RNN, achieve F1 scores of 88.33%,

89.72%, 91.41%, and 93.70%, respectively. IDBN gives the highest F1-score of 89.52%, 91.05%, 92.36%, and 95.15% against MMAS, FFO, BSO, and RCWHO. The IDBN classifier achieves 3.95%, 2.27%, and 0.82% increases in F1-score over GNN, CNN, and RNN.

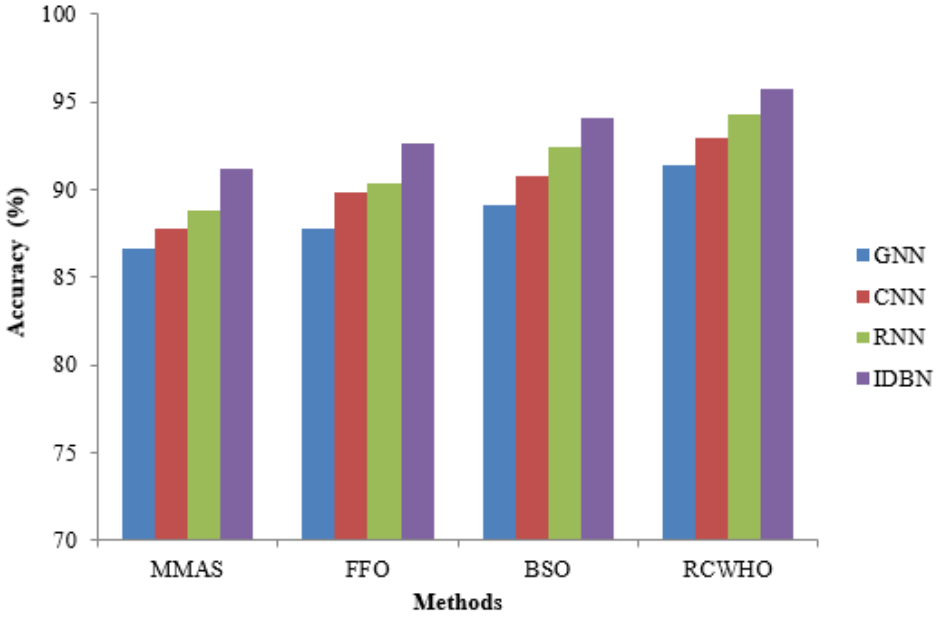

Figure 7: Accuracy analysis comparison of detection methods

Accuracy analysis of classifiers against MMAS, FFO, BSO, and RCWHO is illustrated in Figure 7. GNN achieves accuracies of 86.67%, 87.79%, 89.07%, and 91.39% against MMAS, FFO, BSO, and RCWHO, and CNN achieves accuracies of 87.79%, 89.78%, 90.78%, and 92.95% against MMAS, FFO, BSO, and RCWHO. MMAS, FFO, BSO, and RCWHO with respect to RNN achieve accuracies of 98.84%, 90.35%, 92.41%, and 94.28%, respectively. IDBN achieves the highest accuracies of 8991.17%, 92.58%, 94.05%, and 95.69% against MMAS, FFO, BSO, and RCWHO, respectively. The IDBN classifier achieves 4.3%, 2.74%, and 1.41% higher accuracy than GNN, CNN, and RNN, respectively.

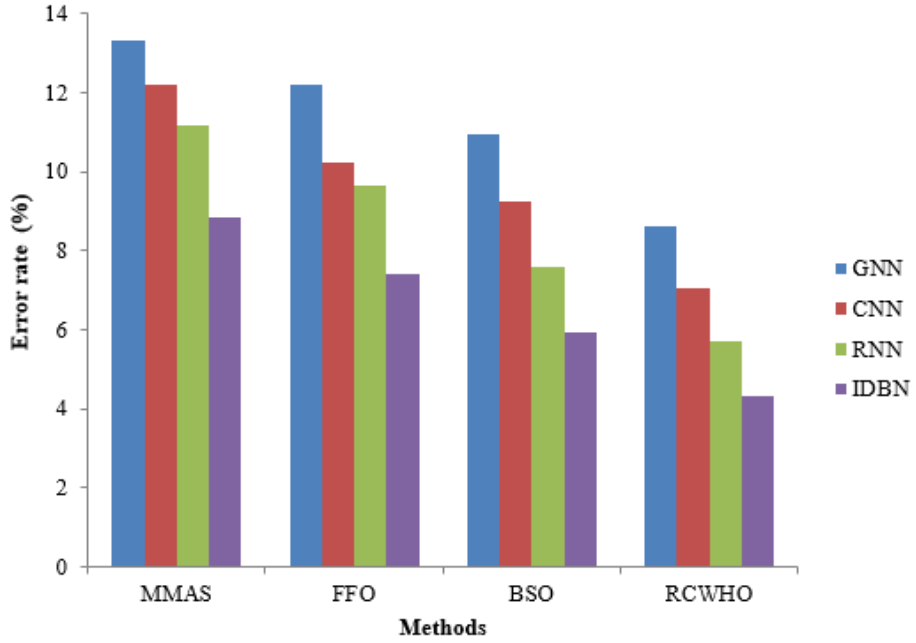

Figure 8: Error rate analysis comparison of detection methods

Figure 8 shows the error rate of classifiers against MMAS, FFO, BSO, and RCWHO. GNN gives error rates of 13.33%, 12.21%, 10.93%, and 8.61% against MMAS, FFO, BSO, and RCWHO, respectively, while CNN gives error rates of 12.21%, 10.22%, 9.22%, and 7.05% against MMAS, FFO, BSO, and RCWHO, respectively. Similarly, the RNN yields error rates of 11.16%, 9.65%, 7.59%, and 5.72% for MMAS, FFO, BSO, and RCWHO, respectively. MMAS, FFO, BSO, and RCWHO, with respect to IDBN, give the lowest error rates of 8.83%, 7.42%, 5.95%, and 4.31%, respectively. The IDBN classifier has the lowest error rates of 4.3%, 2.74%, and 1.41% compared to GNN, CNN, and RNN (Table 1).

Table 1: Evaluation metrics of detection methods

Methods	**Precision (%)**			
	GNN	**CNN**	**RNN**	**IDBN**
MMAS	87.69	89.65	90.88	92.27
FFO	89.24	91.07	92.46	93.82
BSO	90.51	92.18	93.37	94.50
RCWHO	91.93	93.85	95.81	96.45
Methods	**Recall (%)**			
	GNN	**CNN**	**RNN**	**IDBN**
MMAS	85.47	87.03	88.19	90.39
FFO	87.35	88.41	89.67	91.41
BSO	89.08	90.64	91.38	92.67
RCWHO	92.11	93.56	94.49	95.50
Methods	**F1-score (%)**			
	GNN	**CNN**	**RNN**	**IDBN**
MMAS	86.57	88.33	89.52	91.32
FFO	88.28	89.72	91.05	92.60
BSO	89.79	91.41	92.36	93.58
RCWHO	92.02	93.70	95.15	95.97
Methods	**Accuracy (%)**			
	GNN	**CNN**	**RNN**	**IDBN**
MMAS	86.67	87.79	88.84	91.17
FFO	87.79	89.78	90.35	92.58
BSO	89.07	90.78	92.41	94.05
RCWHO	91.39	92.95	94.28	95.69
Methods	**Error (%)**			
	GNN	**CNN**	**RNN**	**IDBN**
MMAS	13.33	12.21	11.16	8.83
FFO	12.21	10.22	9.65	7.42
BSO	10.93	9.22	7.59	5.95
RCWHO	8.61	7.05	5.72	4.31

4.3. Discussion and Findings

CDs are a major problem in the healthcare industry worldwide. The health statement states that the death rate among people rises as a result of CD. Therefore, reducing the patient's risk factors for death is crucial. Computer-Aided Design (CAD) systems are efficient systems for diagnosis. RCWHO has been introduced for feature selection based on an RRA and the CWM. This feature selection approach selects the features based on RRA and CWM with accuracy as the fitness function. Once the features are selected, ISMOTE-based data balancing is proposed for the data imbalance problem. IDBN is introduced for CD prediction. IDBN is assessed using metrics such as precision, recall, F1-score, accuracy, and error rate, with scores of 96.45%, 95.50%, 95.97%, 95.69%, and 4.31%, respectively. The IDBN classifier achieves the highest precision of 92.27%, 93.82%, 94.50%, and 96.45% for MMAS, FFO, BSO, and RCWHO, respectively. The IDBN classifier achieves the highest recall of 90.39%, 91.41%, 92.67%, and 95.50% for MMAS, FFO, BSO, and RCWHO, respectively.

The IDBN classifier achieves the highest F1-Scores of 91.32%, 92.60%, 93.58%, and 95.97% for MMAS, FFO, BSO, and RCWHO, respectively. The IDBN classifier achieves the highest accuracies of 91.17%, 92.58%, 94.05%, and 95.69% for MMAS, FFO, BSO, and RCWHO, respectively. The IDBN classifier has the lowest errors of 8.83%, 7.42%, 5.95%, and 4.31% for MMAS, FFO, BSO, and RCWHO, respectively. The IDBN classifier achieves 4.3%, 2.74%, and 1.41% higher accuracy than GAN, CNN, and RNN, respectively. RCWHO has the highest precision of 91.93%, 93.85%, 95.81%, and 96.45% for GNN, CNN, RNN, and IDBN. RCWHO has the highest recall of 92.11%, 93.56%, 94.49%, and 95.50% for GNN, CNN, RNN,

and IDBN. RCWHO achieves the highest F1 Scores of 92.02%, 93.70%, 95.15%, and 95.97% for GNN, CNN, RNN, and IDBN, respectively. RCWHO achieves the highest accuracy of 91.39%, 92.95%, 94.28%, and 95.69% for GNN, CNN, RNN, and IDBN, respectively. RCWHO has the lowest errors of 8.61%, 7.05%, 5.72%, and 4.31% for GNN, CNN, RNN, and IDBN, respectively.

5. Conclusion and Future Work

In this paper, a deep learning method is introduced to identify and predict the CD patients. For data preprocessing, missing values are imputed using K-Nearest Neighbours (KNN), and the dataset is normalised using Z-score normalisation. For feature selection, the Random Competition Wild Horse Optimizer (RCWHO) is introduced to remove inappropriate and unnecessary features from the normalized dataset. RCWHO is motivated by the collective actions of horses, such as grazing, dominance, leadership hierarchy, and mating, through Random Running Approach (RRA) and Competition Waterhole Mechanism (CWM). Data balancing: The Improved Synthetic Minority Oversampling Technique (ISMOTE) is an effective method for handling imbalanced datasets and has been used for classification. The Improved Deep Belief Network (IDBN) is a deep probabilistic digraph model consisting of several layers with varying numbers of neurons in the CD dataset. After equilibrium is reached, each layer's variables are sampled using a conditional distribution, and the probability value of the next hidden layer is generated one after the other. IDBN is assessed using metrics such as precision, recall, F1-score, accuracy, and error rate, with scores of 96.45%, 95.50%, 95.97%, 95.69%, and 4.31%, respectively. This work has been extended to include an ensemble classifier and an ensemble feature selection model to enhance the accuracy of the present model. In addition, the present system is also extended to real-time datasets with accurate target treatment for every person.

Acknowledgment: The authors sincerely acknowledge the academic support and research facilities provided by Vels Institute of Science, Technology and Advanced Studies for facilitating the successful completion of this research work. They also extend their gratitude to the faculty members and institutional authorities for their valuable guidance, encouragement, and continuous support throughout the study.

Data Availability Statement: The data supporting the conclusions of this study are securely maintained by the authors and may be obtained from the corresponding author upon reasonable request, subject to ethical and institutional requirements.

Funding Statement: The authors collectively confirm that this research was conducted without receiving any external financial assistance, sponsorship, or specific funding support.

Conflicts of Interest Statement: All authors declare that there are no financial, personal, academic, or professional conflicts of interest that could have affected the design, analysis, interpretation, or publication of this research work.

Ethics and Consent Statement: The authors affirm that the study was conducted in accordance with recognized ethical principles, ensuring the confidentiality, privacy, and protection of all participants. Informed written consent was obtained from participants after providing complete details of the study, and all collected information was anonymized and handled in accordance with applicable data protection and ethical guidelines.

References

1. G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," *Journal of Personalized Medicine*, vol. 10, no. 2, p. 21, 2020.
2. B. Manjulatha and P. Suresh, "An ensemble model for predicting chronic diseases using machine learning algorithms," in Smart Computing Techniques and Applications, *Springer*, New York, United States of America, 2021.
3. R. Alanazi, "Identification and prediction of chronic diseases using machine learning approach," *Journal of Healthcare Engineering*, vol. 2022, no. 1, pp. 1–9, 2022.
4. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, no. 4, pp. 8869–8879, 2017.
5. R. Ge, R. Zhang, and P. Wang, "Prediction of chronic diseases with multi-label neural network," *IEEE Access*, vol. 8, no. 7, pp. 138210–138216, 2020.
6. Z. Elkoshi, "The binary classification of chronic diseases," *Journal of Inflammation Research*, vol. 12, no. 12, pp. 319–333, 2019.
7. A. M. Alhassan and W. M. N. W. Zainon, "Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis," *IEEE Access*, vol. 9, no. 6, pp. 87310–87317, 2021.
8. D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informatics Journal,* vol. 19, no. 3, pp. 179–189, 2018.

9. K. Chen, F. Y. Zhou, and X. F. Yuan, "Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection," *Expert Systems with Applications*, vol. 128, no. 8, pp. 140–156, 2019.
10. M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7839–7857, 2020.
11. O. O. Akinola, A. E. Ezugwu, J. O. Agushaka, R. A. Zitar, and L. Abualigah, "Multiclass feature selection with metaheuristic optimization algorithms: A review," *Neural Computing and Applications*, vol. 34, no. 8, pp. 19751–19790, 2022.
12. Z. Sadeghian, E. Akbari, H. Nematzadeh, and H. Motameni, "A review of feature selection methods based on meta-heuristic algorithms," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 37, no. 1, pp. 1–51, 2025.
13. E. S. M. El-Kenawy, S. Mirjalili, F. Alassery, Y. D. Zhang, M. M. Eid, S. Y. El-Mashad, B. A. Aloyaydi, A. Ibrahim, and A. A. Abdelhamid, "Novel meta-heuristic algorithm for feature selection, unconstrained functions and engineering problems," *IEEE Access*, vol. 10, no. 4, pp. 40536–40555, 2022.
14. S. Saturi, "Review on machine learning techniques for medical data classification and disease diagnosis," *Regenerative Engineering and Translational Medicine*, vol. 9, no. 8, pp. 141–164, 2023.
15. P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, no. 4, pp. S36–S40, 2017.
16. K. W. Johnson, J. T. Soto, B. S. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley, and J. T. Dudley, "Artificial intelligence in cardiology," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2668–2679, 2018.
17. C. Zhang, K. C. Tan, H. Li, and G. S. Hong, "A cost-sensitive deep belief network for imbalanced classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 109–122, 2018.
18. H. Lu and S. Uddin, "A weighted patient network-based framework for predicting chronic diseases using graph neural networks," *Scientific Reports*, vol. 11, no. 11, pp. 1–12, 2021.
19. X. Zhang, H. Zhao, S. Zhang, and R. Li, "A novel deep neural network model for multi-label chronic disease prediction," *Frontiers in Genetics*, vol. 10, no. 4, pp. 351–361, 2019.
20. D. Jain and V. Singh, "A two-phase hybrid approach using feature selection and adaptive SVM for chronic disease classification," *International Journal of Computers and Applications*, vol. 43, no. 6, pp. 524–536, 2021.
21. J. J. Gabriel and L. J. Anbarasi, "Optimizing coronary artery disease diagnosis: A heuristic approach using robust data preprocessing and automated hyperparameter tuning of eXtreme gradient boosting," *IEEE Access,* vol. 11, no. 10, pp. 112988–113007, 2023.
22. S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, "Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier," *Applied Sciences*, vol. 10, no. 22, p. 8137, 2020.
23. J. R. Lambert and E. Perumal, "Oppositional firefly optimization based optimal feature selection in chronic kidney disease classification using deep neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 9, pp. 1799–1810, 2022.
24. H. A. Al-Jamimi, "Synergistic feature engineering and ensemble learning for early chronic disease prediction," *IEEE Access*, vol. 12, no. 4, pp. 62215–62233, 2024.
25. A. Singh, N. Prakash, and A. Jain, "Particle swarm optimization-based random forest framework for the classification of chronic diseases," *IEEE Access*, vol. 11, no. 11, pp. 133931–133946, 2023.
26. X. Yuan, S. Chen, C. Sun, and L. Yuwen, "A novel early diagnostic framework for chronic diseases with class imbalance," *Scientific Reports,* vol. 12, no. 5, pp. 1–16, 2022.
27. R. Daid, Y. Kumar, A. Gupta, and I. Kaur, "An effective mechanism for early chronic illness detection using multilayer convolution deep learning predictive modelling," in Proc. *2021 International Conference on Technological Advancements and Innovations (ICTAI),* Tashkent, Uzbekistan, 2021.
28. S. Mishra, H. K. Thakkar, P. Singh, and G. Sharma, "A decisive metaheuristic attribute selector enabled combined unsupervised–supervised model for chronic disease risk assessment," *Computational Intelligence and Neuroscience,* vol. 2022, no. 1, pp. 1–17, 2022.
29. C. H. Cheng, C. P. Chan, and Y. J. Sheu, "A novel purity-based k-nearest neighbors imputation method and its application in financial distress prediction," *Engineering Applications of Artificial Intelligence,* vol. 81, no. 5, pp. 283–299, 2019.
30. S. K. Sehrawat, "Leveraging AI for early detection of chronic diseases through patient data integration," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 3, pp. 125–136, 2024.
31. M. A. Imron and B. Prasetyo, "Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn," *Journal of Soft Computing Exploration,* vol. 1, no. 1, pp. 56–62, 2020.
32. A. A. Ewees, F. H. Ismail, and R. M. Ghoniem, "Wild horse optimizer-based spiral updating for feature selection," *IEEE Access*, vol. 10, no. 10, pp. 106258–106274, 2022.
33. M. P. William, S. B. Kiruba, M. Sakthivanitha, E. S. Soji, and G. Arun, "Machine learning to discover cardiovascular disease onset and key contributors: Data-driven personalized healthcare and preventive strategy," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 3, pp. 137–157, 2024.

34. R. Zheng, A. G. Hussien, H. M. Jia, L. Abualigah, S. Wang, and D. Wu, "An improved wild horse optimizer for solving optimization problems," *Mathematics*, vol. 10, no. 8, p. 1311, 2022.
35. C. S. Kumar, B. S. Kumar, G. Gnanaguru, V. Jayalakshmi, S. S. Rajest, and B. Senapati, "Augmenting chronic kidney disease diagnosis with support vector machines for improved classifier accuracy," in Advances in Medical Technologies and Clinical Practice, *IGI Global*, United States of America, 2024.
36. Y. Bao and S. Yang, "Two novel SMOTE methods for solving imbalanced classification problems," *IEEE Access*, vol. 11, no. 1, pp. 5816–5823, 2023.
37. V. Chunduri, S. S. Rajest, V. R. Allugunti, D. Verma, R. Aarthi, and D. K. Sharma, "Deep neural network for brain tumour segmentation using guaranteed time slots (GTS) algorithm," in Advances in Computer and Electrical Engineering, *IGI Global*, United States of America, 2024.
38. D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, 2023.
39. Z. Zhao, B. Fan, and Y. Zhou, "An efficient water quality prediction and assessment method based on the improved deep belief network—long short-term memory model," *Water*, vol. 16, no. 10, p. 1362, 2024.
40. Kaggle.com. https://www.kaggle.com/datasets/khushikyad001/chronic-disease-progression-tracker-dataset [Accessed by 15/01/2026].