A NOVEL DENSITY-BASED CLUSTERING ALGORITHM FOR PREDICTING CARDIOVASCULAR DISEASE

M. SINDHU

Research Scholar, Department of Computer Science, Gobi Arts and Science College (Autonomous), Gobichettipalayam.

G. T. PRABAVATHI

Associate Professor, Department of Computer Science, Gobi Arts and Science College (Autonomous), Gobichettipalayam.

Abstract

Cardiovascular diseases (CVDs) remain a leading cause of global morbidity and mortality. Early identification of individuals at risk of heart disease is crucial for effective preventive interventions. To improve the prediction accuracy, this paper proposed Heart Disease Prediction using the Density-Based Ordering of Clustering Objects (DBOCO) framework. The Dataset has been pre-processed using Weighted Transform K-Means Clustering (WTKMC). Features are selected using Ensemble Feature Selection (EFS) with a Weighted Binary Bat Algorithm (WBBAT) used to ensure that the emphasis is on the most relevant predictors. Finally, the prediction has been done using the Density-Based Ordering of Clustering method, which has been designed exclusively for cardiovascular disease prediction. DBOCO, a density-based clustering approach, effectively finds dense clusters within data, allowing for the inherent overlap in cardiovascular risk variables. DBOCO captures complicated patterns by detecting these overlapping clusters, improving the accuracy of disease prediction models. The proposed approach has been verified with heart disease datasets, displaying higher performance than traditional methods. This study marks a substantial leap in predicting cardiovascular disease providing a comprehensive and dependable framework for early identification and preventive concern.

Keywords: Cardiovascular Disease, Density-Based Clustering, DBOCO, WBBAT, WTKMC

1. INTRODUCTION

Heart disease datasets can have many features, which can introduce noise, redundancy, and overfitting. Dataset Preprocessing has an important role in machine learning. Overall, feature selection and preprocessing techniques play a crucial role in heart disease prediction by enhancing model performance, reducing dimensionality, improving interpretability, and providing meaningful insights for preventive measures and treatments [27-30]. Bat Algorithm is a powerful optimization technique that can be applied to a wide range of problems. Its ability to handle global optimization, balance exploration and exploitation, and adapt to dynamic environments makes it useful in various fields, including engineering, machine learning, and data science. Prabavathi et al. used Weighted Transform K-means Clustering (WTKMC) an unsupervised machine learning algorithm for partitioning unlabeled data into a defined number of disjoint groups of equal variances based on the weighted transform method. The primary objective of this work is to address the challenges in cardiovascular disease by applying WTKMC for data pre-processing, leveraging Elastic Net with Random Forest Classifier and the Weighted Binary BAT Algorithm for Ensemble Feature Selection [16-20]. The aim is to enhance the

robustness and interpretability of the predictive model. This paper incorporates the Density-Based Ordering of Clustering Objects (DBOCO) algorithm for prediction, providing a comprehensive approach that integrates clustering techniques with ensemble methods.

The remainder of this paper is organized as follows. Research works based on density-based clustering are listed in section 2. Section 3 depicts the proposed model. Section 4 summarizes the results and discussion of the proposed model. Section 5 discusses the conclusion and future work.

2. BACKGROUND STUDY

Clustering is a technique used in machine learning and data analysis to group similar items or data points together based on certain characteristics or features they possess. The goal of clustering is to identify inherent patterns or structures within the data without prior knowledge of the groupings [27].

The partitional clustering algorithm constructs a partition of a dataset D of n objects into a set of k clusters. k is the input parameter which is unfortunately not available for many applications. Normally a partitional algorithm starts with initial partition D followed by an iterative control procedure to optimize an objective function. Hierarchical clustering creates a hierarchical decomposition of dendrogram D. These algorithms iteratively split D into smaller subsets until each subset contains only one object. The advantage of hierarchical clustering is there is no need to specify the number of clusters to be created, but the main problem is the difficulty of deriving parameters for the termination condition [31]. The density-based clustering method connects highly dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. The algorithm connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas. Clustering Large Applications based upon RANdomized Search (CLARANS) is an efficient medoid-based clustering algorithm. In CLARANS, the process of finding k medoids from n objects is viewed abstractly as searching through a graph, where a node is represented by a set of k objects as selected medoids. Two nodes are neighbors if their sets differ by only one object. In each iteration, CLARANS considers a set of randomly chosen neighbor nodes as candidates of new medoids. If the neighbour is a better choice for medoids the algorithm moves to the neighbor. Otherwise, a local optimum is discovered. The entire process is repeated multiple times [26].

Density-based spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm to discover clusters of arbitrary shapes. It requires only one input parameter to determine the appropriate value for it. It starts with arbitrary point p and retrieves all density reachable points from p wrt. Eps (epsilon) and MinPts (minimum points). If p is a core point, DBSCAN forms a cluster wrt. Eps and MinPts. If p is a border point, no points are density reachable from p. So, the algorithm moves to the next point of the database. DBSCAN merges two clusters into one cluster if two clusters have

different densities and are close to each other. DBSCAN outperforms CLARANS by a factor of at least 100 in terms of efficiency [24].

Fitriyani et al. (2020) presented an effective heart disease prediction model for a clinical decision support system, emphasizing the practical implications of predictive models in healthcare settings. Fitriyani et al. (2022) explored a chronic disease prediction model integrating DBSCAN, SMOTE-ENN, and Random Forest, reflecting a trend toward combining clustering techniques with ensemble methods for robust predictions.

Dileep et al., proposed a framework that includes unsupervised learning followed by supervised learning by using Naïve Bayes with DBSCAN achieved 97 % accuracy. One limitation of using machine learning techniques for heart disease prediction is the challenge of working with small medical research datasets. Yu Ping Chang et al., proposed prediction refinement patient adaptation procedure was applied by DBSCAN, HDBSCAN, and OPTICS, effectively adjusting the decision boundary to snap around feature clusters specific to each patient, resulting in improved classification performance. The overall accuracy of the proposed framework is 98.6%.

Samira Ghodratnama et al., proposed an approach for the simplification of complex patterns clustered by various clustering algorithms but the DENCLUE algorithm was selected for the approach which is not sensitive to the input dimension with high dimensional data. In this approach, the author introduces the multimodal distribution to handle datasets and applies the feature selection for each cluster and it produces better results.

2.1 DBSCAN Cluster Model

The model introduced by DBSCAN uses a simple minimum density level estimation, based on a threshold for the number of neighbors, minPts, within the radius ε (with an arbitrary distance measure). Objects with more than minPts neighbors within this radius (including the query point) are a core point. The intuition of DBSCAN is to find those areas that satisfy this minimum density, and which are separated by areas of lower density. For efficiency reasons, DBSCAN does not perform density estimation in between points. Instead, all neighbors within the ε radius of a core point are part of the same cluster as the core point (called direct density reachable). If any of these neighbors is again a core point, their neighborhoods are transitively included (density reachable). Non-core points in this set are called border points, and all points within the same set are density-connected. Points that are not density reachable from any core point are considered noise and do not belong to any cluster. A few variants of DBSCAN that focus on finding hierarchical clustering are OPTICS and HDBSCAN. In HDBSCAN the concept of border points was abundant and only the core points were considered as part of a cluster at any time. This research work uses DBOCO.
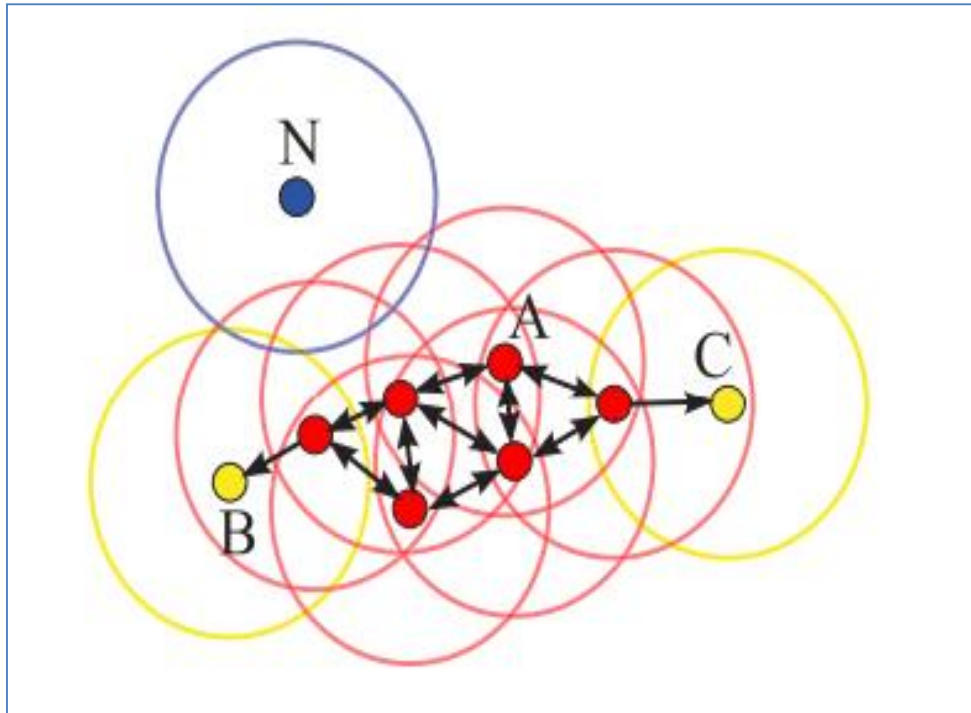
Cardiovascular disease prediction using machine learning and clustering techniques faces several challenges. Integration of multiple algorithms, demographic variability, complexity in feature selection, interpretability of unsupervised techniques, and practical implementation in clinical settings pose significant hurdles. Additionally, optimizing the deployment of ensemble models, balancing computational efficiency with accuracy in hybrid classifiers, and determining the most suitable clustering techniques for CVD prediction remain key challenges. Bridging these gaps is crucial for advancing the field, ensuring the applicability and effectiveness of predictive models in real-world healthcare scenarios, and ultimately contributing to improved patient outcomes and preventive healthcare strategies. In this research work, we suggest DBOCO hierarchical clustering for effective heart disease prediction.

3. MATERIALS AND METHODS

The proposed model described in this paper is to address the problems in cardiovascular disease prediction. The dataset is pre-processed using WTKMC to make it better for analysis. Ensemble Feature Selection uses the WBBAT Algorithm to optimize relevant feature selection for model efficiency and interpretability Prabavathi et al. The DBOCO method is proposed to give a comprehensive predictive framework. DBOCO's density-based cluster ordering allows sophisticated analysis of cardiovascular data hierarchical risk structures. The DBOCO model flowchart is represented in figure 1.
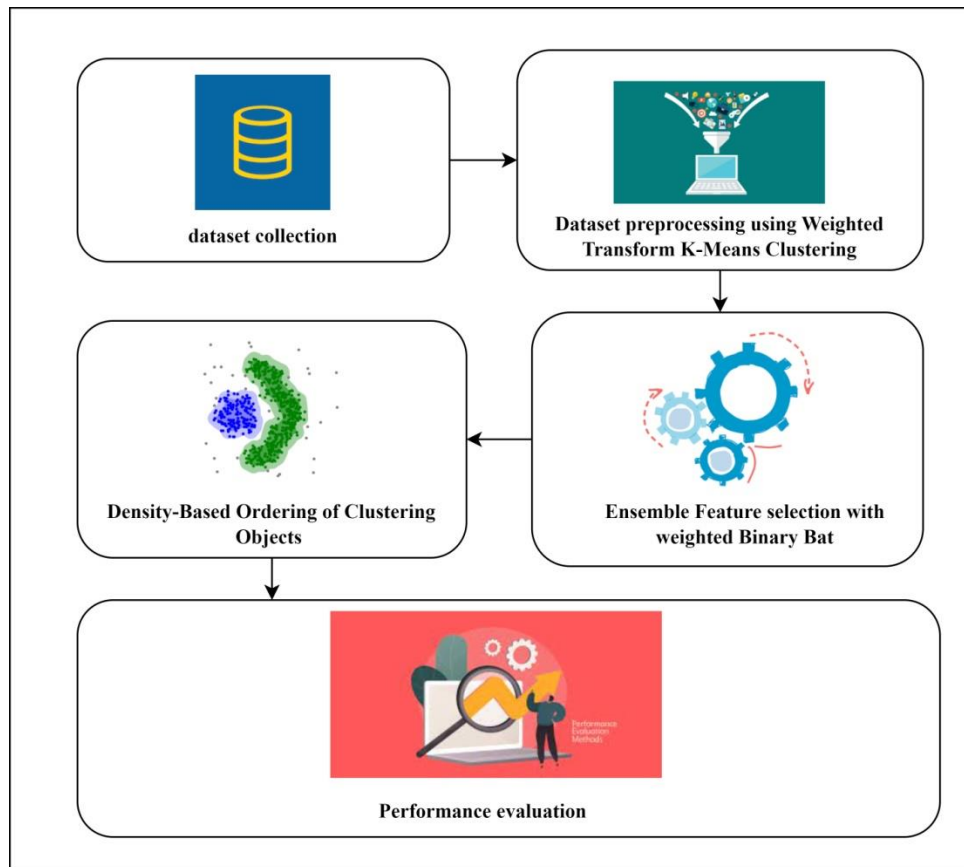
Figure 1: DBOCO Model

3.1 Dataset collection

The dataset has been collected from
https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset   containing   1026
records of patients aged 40-79 years who underwent medical examination. The dataset
includes various patient characteristics such as age, gender, height, weight, systolic and
diastolic blood pressure, and various medical test results such as cholesterol and glucose
levels. It also includes a binary target variable indicating the presence or absence of
cardiovascular disease in each patient and it is commonly used for predictive modelling
tasks.

3.2 Clustering using Density-Based Ordering of Clustering Objects

After Feature selection, when exploring density-based clustering techniques, the DBOCO
emerges as an innovative approach. In the context of heart disease prediction utilizing a
dataset encompassing attributes such as age, gender, cholesterol (chol), resting blood
pressure (thalag), blood pressure (bp), and glucose levels, DBOCO offers a dynamic
adaptation to varying data point densities. This becomes particularly relevant in
cardiovascular datasets where the attributes may contribute to irregularly shaped clusters
with diverse densities. DBOCO excels in detecting clusters that might represent distinct

risk profiles by leveraging the density of nearby data points, a crucial criterion for cluster identification. By employing a density-based ordering technique, DBOCO clusters data points into denser regions, considering attributes like age, gender, chol, thalag, bp, and glucose. This adaptability enables DBOCO to effectively capture complex relationships within the data, contributing to a more precise and versatile clustering process. The density-based clustering algorithm DBSCAN suggested by Z. Xue and H. Wang (2021), has demonstrated effectiveness in handling datasets related to disease prediction. Its ability to efficiently identify clusters, irrespective of their shapes and sizes, aligns with the intricacies of cardiovascular datasets. Nevertheless, challenges in parameter tuning, especially for features like age, gender, chol, thalag, bp, and glucose, remain pertinent. To address this, ongoing research proposes adaptive methods for parameter estimation, enhancing DBOCO's compatibility with high-dimensional datasets and those featuring clusters of varying densities, a crucial consideration when dealing with cardiovascular data. The algorithm has a detailed representation of clustering the dataset for predictive analysis.

Let *db* be a data store holding n items in *d* dimensions. N_ε^{Ai} is a dimensional space, *Ai is the axis*, S is a subset of A denoted by $\forall p \in db$. The notation |*S*| denotes the number of dimensions in the subspace *S*. A density-based clustering technique finds clusters in a database by analyzing the density connection of the data items included within. It looks for density connection by looking at several different items.

Primary Objects: Let $N_\varepsilon^{Ai}(p)$ and $\forall p \in db$, when there is at least l other object within e units of *p* along axis Ai, we refer to p as a core object or a dense object.

$$\forall p \in db, \left| N_\varepsilon^{Ai}(p) \right| \geq \mu \qquad \text{------ (1)}$$

where p's e-neighborhood along dimension Ai is denoted by $\left| N_\varepsilon^{Ai}(p) \right|$. It may be broken down into

$$\left| N_\varepsilon^{Ai}(p) \right| = \{0 \in db | dist(o,p)_{Ai} \leq \varepsilon\} \qquad \text{------- (2)}$$

Here, the distance function along dimension *Ai* between any two objects o and p is denoted as $dist(o,p)_{Ai.}$ The contents of a cluster are defined by the central items that share a shared neighborhood. There are core items and boundary objects in a density-based cluster. An object with fewer than l neighbor in its e-neighborhood is considered as border object. Therefore, the boundary object can be derived by,

$$\forall p \in db, \left| N_\varepsilon^{Ai}(p) \right| < \mu \qquad \text{------ (3)}$$

There are many fewer items in the e-neighborhood of border points than there would be at the density threshold of l. Through density accessible items and density connected objects, the border objects and core objects are linked to one another.

Algorithm: Density-Based Ordering of Clustering
Input:

Input:

- Normalized heart disease dataset X of n data points

- A distance metric d between data points

- A minimum number of points m and a radius ε to define a cluster.

- A threshold distance ρ for connectivity

Steps:

1. Calculate the distance matrix $D = (d_{ij})$ for all pairs of data points in X.

2. For each data point x_i in X, calculate its ε-neighborhood N_i, i.e., the set of all points within distance ε of x_i.

3. For each data point x_i in X, calculate its density ρ_i, i.e., the number of points in its ε-neighborhood: $\rho_i = |N_i|$

4. For each data point x_i in X, find its nearest neighbor x_j with a higher density:

   $N_{N(i)} = \text{argmax} \{j \in N_i \cup \{i\}\} \{\rho_j\}$ subject to $d(i, j) \leq \rho_i$

5. Define a reachability distance for each pair of points (x_i, x_j):

   $RD(x_i, x_j) = \max\{\rho_i, d(i, j)\}$

6. Sort the data points in decreasing order of density: $\rho_1 \geq \rho_2 \geq ... \geq \rho_n$

7. Initialize a priority queue Q and a current cluster $C = \emptyset$

8. For each data point x_i in the sorted order:

   a. If x_i has not yet been processed, add it to C and to Q with priority ρ_i.

   b. While Q is not empty, remove the highest-priority point x_j from Q.

       i. If x_j has not yet been processed, add it to C and to Q with priority ρ_j.

       ii. For each unprocessed point x_k in the ε-neighborhood of x_j:

           1. Calculate the new reachability distance $RD'(x_k) = \max\{RD(x_j, x_k), \rho_j\}$.

           2. If $RD'(x_k) \leq \rho_k$ and x_k has not yet been added to Q, add it to Q with priority $RD'(x_k)$.

       iii. Mark x_j as processed.

   c. If C contains at least m points, output the center of C as a cluster center.

Stop when all data points have been processed.

Output: Prediction Accuracy

Time complexity refers to the computational efficiency of the various steps involved in this pipeline. Time complexity is often represented using the "Big O" notation, denoted as $O(f(n))$, where 'f(n)' describes the upper bound on the growth rate of the execution time concerning the input size 'n'.

$$O(f(n)) \qquad \qquad \text{--------- (4)}$$

Here, 'f(n)' represents the time complexity associated with data preprocessing operations. The specific value of 'f(n)' will depend on the operations performed and the dataset size.

4. RESULTS AND DISCUSSION

The study aims to identify the most efficient model for diagnosing CVD based on the patient's health history. In this section, we present the DBOCO model outcomes of our heart disease prediction model leveraging the advanced clustering approach of DBOCO. The proposed model incorporates essential attributes such as age, gender, cholesterol (chol), resting blood pressure (thalag), blood pressure (bp), and glucose levels for a comprehensive analysis. The clustering output is shown in figure 2.
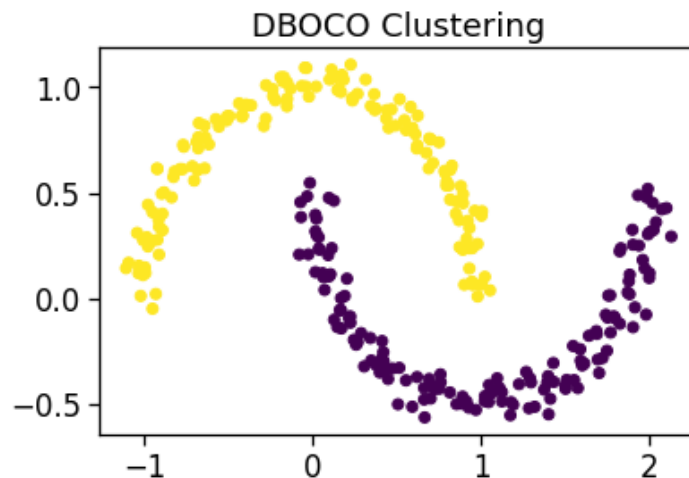

Figure 2: DBOCO Clustering

The algorithms used for comparison are DENCLUE, DBSCAN, HDBSCAN, OPTICS, Naïve Bayes with DBSCAN, and DBOCO. For 1026 records, the number of clusters varied between 1 and 2 and the cluster value count ranged from 10 to 14.

Figure 3 provides valuable insights into the time complexity of various clustering and machine learning methods across distinct stages of their execution. In terms of pre-processing time, DBOCO emerges as the swiftest, followed by DENCLUE, OPTICS, Naïve Bayes with DBSCAN and HDBSCAN. However, when it comes to the actual clustering process, DBOCO stands out as the most time-consuming, with Naïve Bayes with DBSCAN, HDBSCAN, DBSCAN, and OPTICS following in order. For train-test data splitting, HDBSCAN requires the most time, while Naïve Bayes with DBSCAN, OPTICS, DBSCAN, and DBOCO exhibit lower time complexities. Finally, during the training phase,

HDBSCAN demands the longest time, followed by Naïve Bayes with DBSCAN, OPTICS, DBSCAN, and DBOCO.
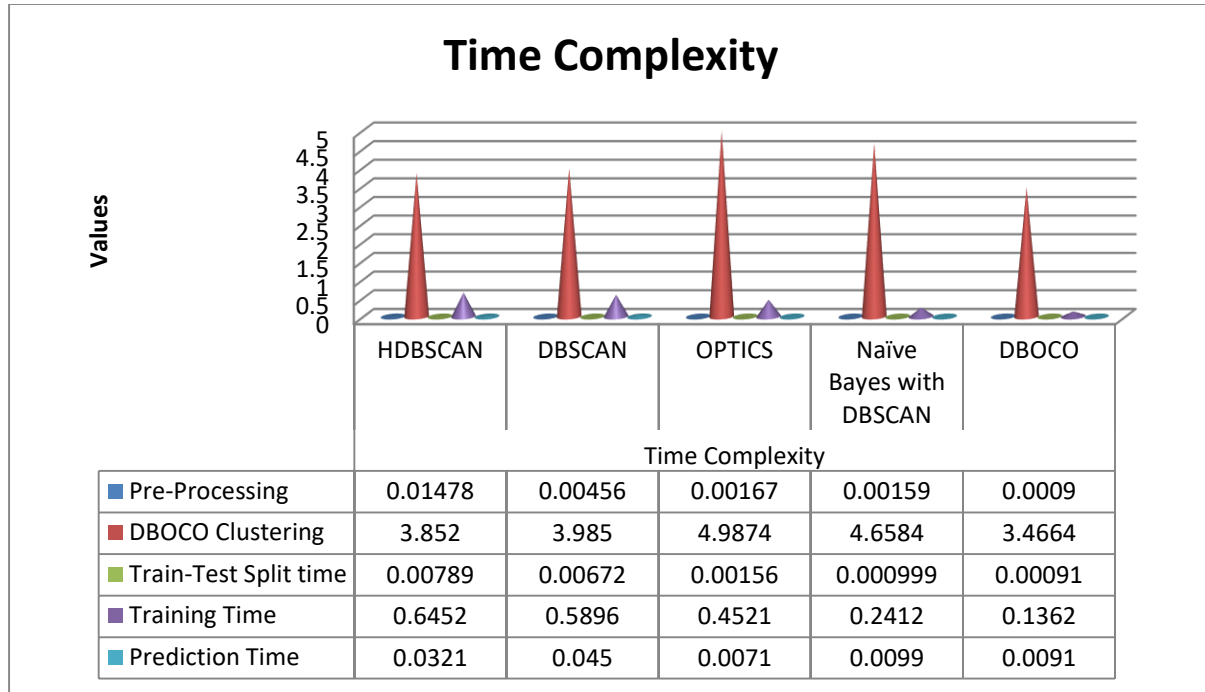

Time Complexity

Time Complexity	HDBSCAN	DBSCAN	OPTICS	Naïve Bayes with DBSCAN	DBOCO
■ Pre-Processing	0.01478	0.00456	0.00167	0.00159	0.0009
■ DBOCO Clustering	3.852	3.985	4.9874	4.6584	3.4664
■ Train-Test Split time	0.00789	0.00672	0.00156	0.000999	0.00091
■ Training Time	0.6452	0.5896	0.4521	0.2412	0.1362
■ Prediction Time	0.0321	0.045	0.0071	0.0099	0.0091

Figure 3: Time Complexity Comparison chart

Table 1 shows performance metrics (Accuracy, Precision, Recall, and F-measure) of different clustering algorithms before and after feature selection are presented. Before feature selection, DENCLUE achieved an accuracy of 89% with corresponding precision, recall, and F-measure of 89%, 90%, and 89% respectively. HDBSCAN demonstrated higher accuracy at 93%, with precision and recall at 91% and 90% respectively, but a slightly lower F-measure of 88%. DBSCAN showed impressive results with 94% accuracy, 91% precision, 92% recall, and 96% F-measure. OPTICS had lower performance metrics, with an accuracy of 49%, precision of 51%, recall of 53%, and F-measure of 52%. Naïve Bayes combined with DBSCAN yielded excellent results, boasting 97% accuracy, 96% precision, 95% recall, and 96% F-measure. DBOCO also performed well with 95% accuracy, 94% precision, 91% recall, and 92% F-measure. After feature selection, DENCLUE accuracy slightly increased to 90%, with improved precision and F-measure at 91% and 89% respectively. HDBSCAN showed a minor boost in accuracy to 94%, with precision and recall at 92% and 92% respectively, and an F-measure of 90%. DBSCAN maintained its performance with 95% accuracy, 95% precision, 94% recall, and 96% F-measure. OPTICS showed improvement in recall (61%) but a decrease in other metrics, resulting in an accuracy of 53%, precision of 50%, and F-measure of 60%. Naïve Bayes with DBSCAN maintained high performance with 96% accuracy, 94% precision, 96% recall, and 91% F-measure. Notably, DBOCO exhibited significant improvement after feature selection, achieving an accuracy of 98.33%, perfect

precision of 100%, recall of 97.03%, and F-measure of 98.49%. Overall, feature selection positively impacted most algorithms, enhancing their performance in the clustering task.

Table 1: Performance metrics

	Methods	Accuracy	Precision	Recall	F-measure
Before Feature Selection	DENCLUE	89	89	90	89
	HDBSCAN	93	91	90	88
	DBSCAN	94	91	92	96
	OPTICS	49	51	53	52
	Naïve Bayes with DBSCAN	97	96	95	96
	DBOCO	95	94	91	92
After Feature Selection	DENCLUE	90	91	90	89
	HDBSCAN	94	92	92	90
	DBSCAN	95	95	94	96
	OPTICS	53	50	61	60
	Naïve Bayes with DBSCAN	96	94	96	91
	DBOCO	98.33	100	97.03	98.49

V. CONCLUSION

The DBOCO model proves to be instrumental in significantly enhancing the accuracy and precision of heart disease prediction, especially when implemented after feature selection. This study uses a comprehensive and refined approach for analyzing intricate datasets, integrating Weighted Transform K-Means Clustering for Preprocessing, Ensemble Feature Selection with the Weighted Binary Bat algorithm for feature selection and clustering using the Density-Based Overlapping Clustering algorithm. The adaptability of DBOCO to diverse data densities and intricate feature relationships positions it as a useful algorithm in the domain of cardiovascular health analytics. The findings suggest that the incorporation of DBOCO, especially in synergy with feature selection, holds considerable promise for advancing the accuracy and reliability of predictive models in cardiovascular disease diagnosis and prognosis. With 98.33% accuracy, DBOCO has the potential to enhance healthcare practices. In furtherance, we will apply the various clustering algorithms to improve accuracy.

Reference

1) Rahim.A, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "*An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases*," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

2) Mohan, S., Thirumalai, C., & Srivastava, G., "*Effective Heart Disease Prediction using Hybrid Machine Learning Techniques*". IEEE Access, 2019, doi:10.1109/access.2019.2923707

3) A. Kumar, K. U. Singh and M. Kumar, "*A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method*," in *Big Data Mining and Analytics*, vol. 6, no. 4, pp. 513-525, December 2023, doi: 10.26599/BDMA.2022.9020052.

4) Fitriyani N.L, M. Syafrudin, G. Alfian and J. Rhee, "*HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System*," in *IEEE Access*, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

5) Prabavathi.G.T, Sindhu.M, "*DNFH: Data Normalization And Feature Selection Using Hybrid Methods For Heart Disease Prediction*", Journal of Namibian Studies - History Politics Culture (ISSN print 1863-5954; ISSN online 2197-5523), Vol. 35 (2023): Special Issue 1, DOI https://doi.org/10.59670/jns.v35i.3775

6) Ashri, M. M. El-Gayar and E. M. El-Daydamony, "*HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm*," in *IEEE Access*, vol. 9, pp. 146797-146809, 2021, doi: 10.1109/ACCESS.2021.3122789.

7) Fitriyani.N.L, M. Syafrudin, G. Alfian, C. -k. Yang, J. Rhee and S. M. Ulyah, "*Chronic Disease Prediction Model Using Integration of DBSCAN, SMOTE-ENN, and Random Forest*," *ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, Manama, Bahrain, 2022, pp. 289-294, doi: 10.1109/ICETSIS55481.2022.9888806.

8) Raihan.M, M. T. Islam, F. Farzana, M. G. M. Raju and H. S. Mondal, "*An Empirical Study to Predict Diabetes Mellitus using K-Means and Hierarchical Clustering Techniques*,"*10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944552.

9) Singh, Reetu & Elangovan, Rajesh.," *Prediction of Heart Disease by Clustering and Classification Techniques*". International Journal of Computer Sciences and Engineering, 2019, doi: 7. 10.26438/ijcse/v7i5.861866.

10) Zhou, Z., Wang, Y., & Li, M., "*Feature Selection Method Based on Hybrid SA-GA and Random Forests*" International Conference on Computing and Data Science (CDS), 2020, doi:10.1109/cds49703.2020.00034

11) Xue. Z., & H. Wang, "Effective density-based clustering algorithms for incomplete data," in Big Data Mining and Analytics, vol. 4, no. 3, pp. 183-194, Sept. 2021, doi: 10.26599/BDMA.2021.9020001.

12) Dileep, P.," Impact *of K-Means and DBSCAN Clustering on Supervised Learning for Heart Disease Prediction*", Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, vol. 12, no. 9, 1173-1183.

13) Mohan, S., Thirumalai, C., & Srivastava, G., "*Effective heart disease prediction using hybrid machine learning techniques*". IEEE Access, 2019, vol. 7, pp. 81542-81554.

14) Chandralekha, M., & Shenbagavadivu, N.," *Performance analysis of various machine learning techniques to predict cardiovascular disease: An empirical study*" Appl. Math. Inf. Sci, 2018, vol. 12 no.1, pp. 217-226.

15) Monia, H. Suyal and A. Gupta, "*An Improved Multi-label k-Nearest Neighbour Algorithm with Prototype Selection using DENCLUE*," *9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2021, pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596427.

16) Xin-She Yang., "*Bat Algorithm: Literature Review and Applications*", Int. J. Bio-Inspired Computation, 2013, vol 5. No. 3, pp 141-149, doi:10.1504/IJBIC.2013.055093.

17) Seyedali Mirjalili., Seyed Mohammad Mirjalili., & Xin-She Yang, "*Binary bat algorithm*", Neural Computing & Applications, Springer 2013, doi: 10.1007/s00521-013-1525-5.

18) Adele Cutler, D. Richard Cutler & John R. Stevens, "*Random Forests*", Ensemble Machine Learning Methods, and Applications, 2011, pp 157-176, doi: 10.1007/978-1-4419-9326-7_5.

19) Bahr kadhim Mohammed & Ashwaq Abdul Sada Kadhim.," Elastic *Net Variable Selection Regularization Method for Regression Discontinuity Designs with Application"*, International Journal of Agricultural and Sciences, 2021, vol. no. 17, pp. 1423-1430, doi:03899.2021.17.1423.

20) Prabavathi G.T. & Sindhu M, "*A Review on Heart Disease Predictive System Using Machine Learning Algorithms"*, Shodha Prabha (UGC CARE Journal), 2022, vol. no. 47, pp 37-42.

21) Tania, F. A., & Shill, P. C., "*A Hybrid Kernal Support Vector Machine with Feature Selection for the Diagnosis of Diseases"*, 4th International Conference on Electrical Information and Communication Technology (EICT), 2019, doi:10.1109/eict48899.2019.9068804.

22) Nanehkaran.Y.A ,1 Zhu Licai., "*Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach"*, Hindawi Wireless Communications and Mobile Computing ,2022, doi:10.1155/2022/6913043

23) Ritesh Sonawane Hitendra D.Patil, " *Prediction of Heart Disease by Optimized Distance and Density-Based Clustering"*, 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, doi: 10.1109/ICAIS53314.2022.9742885

24) Martein Ester, *"A density-based algorithm for discovering clusters"*, KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

25) Martin and Erich, "*DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN"*, ACM Transactions on Database Systems, vol. 42, no 3, pp 1–21, doi: 10.1145/3068335.

26) R. Ng and J. Han., "CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE Trans. Knowledge and Data Engineering, 2002.

27) Driver and Kroeber, "*Quantitative Expression of Cultural Relationships*". University of California Publications in American Archaeology and Ethnology. Berkeley, CA: University of California Press. Quantitative Expression of Cultural Relationships: 211–256, 1932.

28) Zubin, Joseph, "*A technique for measuring like-mindedness*". The Journal of Abnormal and Social Psychology, vol. 33 no.4, pp: 508–516, 1938, doi: 10.1037/h0055441*.*

29) Tryon, Robert C., "*Cluster Analysis: Correlation Profile and Orthometric (factor)"* Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers, 1939.

30) Cattell, R. B., "The description of personality: Basic traits resolved into clusters". Journal of Abnormal and Social Psychology, 1943, vol. 38, no. 4 pp: 476–506, doi: 10.1037/h0054116.

31) Philip K. Hopke, Leonard Kaufman, "*The use of sampling to cluster large data sets*" ELSEVIER Chemometrics and Intelligent Laboratory Systems, vol. 8, no 2, 1990, pp. 195-204.

32) Yu-Ping Chang, Sung-Nien Yu," *Improving Inter-patient Electrocardiogram Diagnosis through Unsupervised Patient Adaptive Feature Density-based Clustering"* ELSEVIER, 2023, doi:10.2139/ssrn.4594340

33) Samira Ghodratnama., & Reza Boostani., "*An Efficient Strategy to Handle Complex Datasets Having Multimodal Distribution*" Springer International Publishing Switzerland, DOI: 10.1007/978-3-319-10759-217.