Pertanika J. Sci. & Technol. 32 (1): 265 - 284 (2024)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Soft Set-based Parameter Reduction Algorithm Through a Discernibility Matrix and the Hybrid Approach for the Risk-Factor Prediction of Cardiovascular Diseases by Various Machine Learning Techniques

Menaga Anbumani[1] and Kannan Kaniyaiah[2]*

[1]*Department of Computer Science and Engineering, Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam-612001 Tamil Nadu, India*
[2]*Department of Mathematics, Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam-612001 Tamil Nadu, India*

ABSTRACT

Parameter reduction without performance degradation is a promising task in decision-making problems. For instance, a great challenge exists in constructing cost functions in gaming theory. Nevertheless, soft set theory handles all its drawbacks conveniently through a new tool for the choice function mathematically. In this paper, we propose an algorithm (SSPRDM) for parameter reduction of soft sets through discernibility matrices, and it is implemented in detecting the risk factor of heart disease problems by using six types of machine learning techniques. The parameters are extracted from the heart disease patient data by the SSPRDM algorithm, and then six machine learning techniques (LDA, KNN, SVM, CART, NB, RF) are performed in the prediction of risk factors for heart disease. The experimental results showed that the present hybrid approach provides an accuracy of 88.46% in the Random Forest technique, whereas the same Random Forest classifier provides an accuracy of 69.23% in the prediction of risk factors of cardiovascular disease (CVD) diagnosis in the earlier work which is a drastic improvement. Moreover, out of 18 parameter reductions, the core component is identified as Total Cholesterol, which is to be considered in all types of CVD diagnosis, whereas Sugar-Fasting (C), Total-Cholesterol (G), and HDL-Cholesterol (I) are the core components identified in three parameter reductions ABCEGHI, ACFGIJ, and BCFGIJK.

Keywords: Discernibility function, discernibility matrix, parameter reduction, risk factor, soft sets

ARTICLE INFO

Article history:
Received: 04 April 2023
Accepted: 17 August 2023
Published: 24 November 2023

DOI: https://doi.org/10.47836/pjst.32.1.16

E-mail addresses:
anbumenakaa@gmail.com (Menaga Anbumani)
anbukkannan@rediffmail.com (Kannan Kaniyaiah)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The difficulties that existed in the dissimilarities of set theory were overcome by the concept of a soft set with suitable parameterizations (Moldostov, 1999). According to Moldostov (1999), the definition of a soft set is a couple (S, A) over (X) iff there exists a mapping from S on A to the collection of all subsets of the set X. Soft sets sidestep the limitations on suitable interpretation in such a way it gets an apt and easy tactics in exercise. There is a challenge in constructing cost function as far as gaming theory is concerned. Nevertheless, soft set theory handles all its drawbacks conveniently through a new tool for the choice function mathematically. The parameter reduction (Xie, 2016) on soft sets has been widely exposed in the history of soft sets in recent days.

The usage of rough sets (Pawlak, 1982; Pawlak, 1991) in decision-making applications is given by Maji et al. (2002). In the interim, some unsuitable and illogical reports in minimizing the parameter are projected (Maji et al. 2002). These irrational reports were pointed out and enhanced to reduce the complexity of collecting information and making better conclusions through soft sets (Chen et al. 2005). Kovkov et al. (2007) used optimization problems to reflect the idea of soft sets and also concentrated on validating the idea of the estimated description of objects. The clue for a standard decrease of parameters in soft sets is taken into account to find solutions to the problems of suboptimal choice and supplemented parameter sets of soft sets (Kong et al., 2008).

The AND operation is used to reduce multi-valued information systems based on soft-set. This strategy is recommended by Herawan et al. (2009). Rose et al. (2010) have developed a novel approach for finding maximal supported sets using parameterizations and diminution. These supported sets have been determined from the Boolean value information system. This proposed methodology is considered to be a reliable decision-making system. An uni-int decision-making method was provided (Cagman & Enginoglu, 2010). This method links the decision function to choose the optimal choice fruitfully in day-to-day problems that have no certainties naturally by descending as wide-ranging replacements.

A novel standard parameter-lessening procedure using soft sets was recommended (Ma et al., 2011). It is grounded on oriented-parameter sum ignoring parameter significant grade and choice partition. A method was proven to diminish the number of parameters for a soft set by not varying its unique classification ability (Ali, 2012). Considering the age factors, prostate volume (PV), and prostate-specific antigen (PSA) of patients, a Soft Expert System (SES) was developed to predict prostate cancer (Yuksel et al., 2013).

Normal parameter reductions based on soft sets are used to design the harmony search algorithm and the intelligent optimization algorithm and are applied to solve the same problems in data mining. By removing the unessential core, Kong et al. (2014) used the particle swarm optimization method to minimize the attributes in the soft set. A technologically progressive HPC algorithm is used to advance the choice of search area,

which is also used to reduce choice costs by streamlining the real categorizations by choice partition order (Mohammed et al., 2017). The restriction in the current prevailing algorithm is broken by Khan and Zhu (2019), who proposed a procedure for standard parameter reduction of soft set that finds application in the medical field. Sadiq et al. (2020) principally focused on lessening implementation expenses by adapting the innovative organizations by choice barrier order and refining the likelihood of probing the domain's realm through an advanced Markov chain model.

In the medical domain, there are many works related to the application of machine learning in forecasting diseases (Mitchell, 1997). The major impact on the heart is due to critical conditions caused by cardiovascular disease (Boukhatem et al., 2022). It requires early disease prediction by finding reliable, accurate, and sensible approaches to identifying the numerous risk factors related to heart (Shah et al., 2020). Jindal et al. (2021) classified the patients with heart disease using logistic regression and KNN algorithms. The prediction of sudden bursts of heart-related diseases among patients on dialysis grounded on machine learning is given accurately by Mezzatesta et al. (2019). Bhat et al. (2022) created a strategy for predicting the risk of developing diabetes in North Kashmir using machine learning algorithms. Six MLA, including Random Forest (RF), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Gradient Boost (GB), Decision Tree (DT), and Logistic Regression (LR), have been used successfully in the experimental study. For the balanced data set, RF has the highest accuracy rate of any classifier.

A method to predict cardiac disease was proposed by Ansari et al. using a variety of machine learning algorithms (MLA), including logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), Naive Bayes (NB), random forest (RF), and decision tree (DT) (Ansari et al., 2023). They assessed the model's performance in predicting heart disease using the testing data set, and their findings indicated that, compared to other algorithms, the k-nearest neighbor method and random forest best fit the data.

In the ground field of medical sciences, these methods were beneficial in examining and estimating a diversity of medicinal syndromes (Alotaibi, 2022). Nonalcoholic fatty liver disease (NAFLD) is common among patients and results in cardiovascular disease (CVD), which acts as the key source of death (Sharma et al., 2022). Palaniappan et al. (2021) developed a machine learning model on modified SEIR, which is used to forecast the epidemic disease dynamics on the contamination risk. The model used Multivariate Logistic Regression. In early 2020, AlArjani et al. (2022) witnessed an outburst of COVID-19, and the entire globe was affected by the same. Machine learning is an important tool in screening, analyzing, forecasting, tracking, and predicting the features and tendencies of COVID-19 (Rahman et al., 2021). Even though many articles on COVID-19 were published in 2020, no effective estimation means still exist to identify the disease with a hundred percent efficiency (Abirami & Kumar, 2022). The most common heart disease is coronary

artery disease (CAD). Over 80% of the deaths are due to this disease, common among developing countries, including Nigeria. This disease's victims are under 70 (Muhammad et al., 2021).

According to research, Alzheimer's disease, a frequent form of dementia, has no known cure. However, by raising the patient's quality of life and offering a solution to strengthen the patient's cognitive capacities with Smartphones, the disease's progression can be slowed down. It highlights the necessity for relatively straightforward cross-platform mobile application development with interactive GUIs to improve users' cognitive capacities (Gupta et al., 2018).

For more effective and accurate recognition of Tetralogy of Fallot (TOF), Wang et al. (2021) suggest a new artificial intelligence model based on cardiovascular computed tomography. Their model combines stochastic pooling, structural optimization, and convolutional neural networks. It is known as SOSPCNN (structurally optimized stochastic pooling convolutional neural network). Multiple-way data augmentation is also performed to avoid overfitting. The proposed SOSPCNN model is given explainability using Grad-CAM. Meanwhile, this SOSPCNN model is used to create desktop and web apps.

Zhang and colleagues used two-level stationary wavelet entropy (SWE) to extract features from brain images. Next, they examined three machine learning-based classifiers: the decision tree, k-nearest neighbors (kNN), and support vector machine (Zhang et al., 2016). According to their experimental findings, the kNN performed the best among the three classifiers. Furthermore, the proposed SWE+kNN technique outperforms four cutting-edge approaches.

It motivates that more work on hybrid approach is essential in the routine life problems. The current article aims to highlight effective procedures for parameter lessening using soft sets with a discernibility matrix and to identify the core parameters involved in diagnosing heart disease. Further, it is utilized to estimate the influence of risk associated with heart ailment using machine learning techniques.

Soft Sets and Information Systems

Consider that X is a determinate collection of objects, and E is a finite set of attributes. The pair (X, E, V, f) is known as an information system if f is an information function from $X \times E$ to $V = \bigcup_{\alpha \in E} V_\alpha$ where every $V_\alpha = \{ f(x, \alpha), \alpha \in E, x \in X \}$ is the value of the attribute α. An information system (X,E, V, f) is called a bi-value if V = {0, 1}. Suppose that $S = f_E$ is a soft set over X. Then I_S = (X,E, V, f_S) is called the bi-value information system persuaded by S where fS: $V \times E \rightarrow V$ is well-defined by $f_S(x, a) = \begin{cases} 1, & x \in F(\alpha) \\ 0, & x \notin F(\alpha) \end{cases}$, for any $x \in X$, $\alpha \in E$. Let I = (X,E, V, f) be a bi-value information system. Then, S_I =

(f_I ,E) is called the soft set over U persuaded by I, where F_I: A → 2^X is well-defined by $F_I(\alpha) = \{x \in X / f(x,\alpha) = 1, \alpha \in E\}$. Suppose that ∑ = {S/ S = f_A is a soft set over U} and Γ = {I/I = (X,E, V, f) is a bi-value information system}. Then, there occurs a one-to-one relation between ∑ and Γ. Let f_E be a soft set over X and let (X, E, V, f) be the bi-value information system persuaded by f_E over X. For any B ⊆ A, the association R_B is stated as follows: R_B = {(x, y) ∈ X × X/f(x, α) = f(y, α), ∀ α ∈ B}. One can straightforwardly witness that $R_{\{\alpha\}} = R_\alpha$ and $R_B = \bigcap_{\alpha \in B} R_\alpha$. Let f_E be a soft set over X. (1) Any B ⊆ A is a parameter lessening of f_E if $R_A = R_B$ and $R_A \neq R_{B-\{\alpha\}}$ for any α ∈ B. (2) The common collection of all parameters lessening f_E is called the softcore and is designated by core(f_E). Suppose f_E is a soft set over |X| = n and let (X, L, V, f) be the bi-value info system persuaded by f_E over X. For x, y ∈ X ,d(x, y) is defined as follows: d(x, y) = {e ∈ E / g(x, e) ≠ f(y, e)}. (1) d(x, y) is the collection of parameters that can distinguish the objects x and y. (2) $D(f_E) = (d_{ij})_{n \times n}$ is called the discernibility matrix of f_E, where X = {x_1, x_2, ..., x_n} and d_{ij} = d (x_i, x_j).

MATERIALS AND METHODS

Vignesh Clinical Laboratory, Karthik Laboratory and Keerthi X-Ray, and Vaishnavi Clinical Laboratory are renowned laboratories in Kumbakonam, Tamilnadu, India. Data is collected from the laboratories mentioned above. The dataset contains details: Blood-Glucose, Blood−Urea, Serum−Creatinine, Total−Cholesterol, HDL−Cholesterol, LDL−Cholesterol, Triglycerides, Risk−Factor−I and Risk−Factor−II. The total number of records is 110. First, some efficient procedures to minimize the parameters using soft sets are proposed. It is then used to predict the estimation of risk factors associated with heart ailment using machine learning techniques. Finally, the results are discussed and compared with those obtained by the authors with standard parameter lessening (without soft sets) and are visualized in Figure 1.

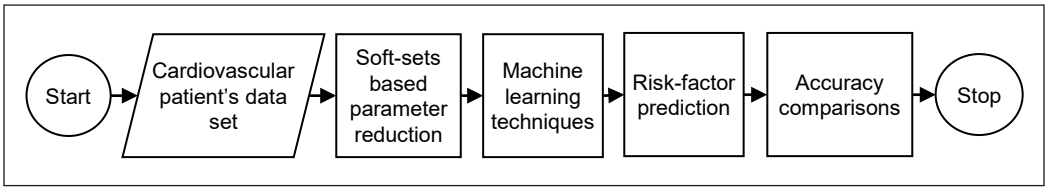


Figure 1. Methodology

Proposed SSPRDM Algorithm for Minimizing the Parameters Using Soft Sets

The succeeding discussion studies the procedure for effective parameter lessening through a discernibility matrix, which contains the resulting steps and applies to decision-making problems.

Step 1: Finding Discernibilty_Matrix

The algorithm Discernibilty_Matrix takes map_matrix[][]as input (Kannan & Menaga, 2022). It constructs the dmatrix of size n×n. Here, 'n′ defines the total number of objects. The dmatrix[i][j] = null for every rowvalue(i) = columnvalue(j). That is, the diagonal entries in the dmatrix are set to null. Each entry in the dmatrix is found by concatenating the features[k] where k = 1 to f provided the condition map_matrix[k][j] not equal to map_matrix[k][i] is satisfied. The process is repeated for n times. In the succeeding algorithm, n is denoted as num.

Algorithm 1 Discernibility_Matrix(map_matrix[][])

```
1: value = empty
2: for x ← 1 to num do
3:    for y ← 1 to num do
4:              for k ← 1 to f do
5:                        if matrix[k, y] <> matrix[k, x] then
6:                                  value = value + features[k]
7:                        end if
8:                        Next k
9:              end for
10:             dmatrix[x, y] = value
11:             value = empty
12:             Next y
13:  end for
14:  Next x
15: end for
16: Display the dmatrix
```

The number of comparisons done gives the time complexity to generate the discernibility_matrix. The inner loop consists of one comparison repeated for n × n × f. Hence, the algorithm's efficiency is given by O (n^3) as f approaches n.

Step 2: Finding Discernibility_Function

The algorithm takes Discernability_matrix[][] as input and finds unique entries such that no two entries are exactly equal. The first step removes the spaces in the leading and trailing ends of the entries using trim(). Then, it takes the entries one by one, compares them with the remaining entries, and sets null for the exactly equal entries. After n × n iterations, the matrix consists of unique parameters and null entries. Excluding null entries, the remaining parameters are stored in the Discenibilityfn_matrix for further processing. Regarding time complexity, each entry is compared with all the remaining entries, and n is the count of the objects. Hence, O (n^2) gives the overall complexity.

Algorithm 2 Discernibility_Fucntion (Discernibility_matrix[][],num)

```
1: for x ← 1 to num do
2:    for y ← 1 to num do
3:              Discernibility_matrix[x][y] = Discernibility_matrix[x][y].Trim
4:              Next y
5:    end for
6:    Next x
7: end for
8: for x ← 1 to n do
9:    for y ← 1 to n do
10:             S = Discernibility_matrix[x][y]
11:             for h ← 1 to num do
12:                       for k ← 1 to num do
13:                                 if S.Equals(Discernibility_matrix[h][k]) then
14:                                           Discernibility_matrix[h][k]=“”
15:                                 end if
16:                                 Next k
17:                       end for
18:                       Next h
19:             end for
20:             Next y
21:  end for
22:  Next x
23: end for
24: q = 1
25: for x ← 1 to num do
26:  for y ← 1 to num do
27:             if (Discernibility_matrix[i][j] <> “”) then
28:                       Discernibilityfn_matrix[q] = Discernibility_matrix[h][k]
29:                       q = q + 1
30:             end if
31:             Next y
32:  end for
33:  Next x
34: end for
```

Step 3: Parameter Reduction_Discernibility_function

The chief goal of finding discernibilityfn_matrix is to find the parameter reduction from the discernibility matrix. It consists of unique entries. The next step is to take each parameter and to find the list of parameters containing it, and it can be concluded that all the parameters

containing it can be reduced to the parameter taken for the substring comparison. For this, the Instr() function returns the index of the first occurrence of the substring contained in the original string. Nevertheless, Instr() finds the pattern in another string bit, and it does not return true if the original string contains the substring but not in the exact pattern. Another function named check is used to find that.

The check function takes two parameter entries as strings as input. It finds the length of the two strings. The function also converts the two strings to a character array. It compares and finds the minimum length between the two strings. It then checks whether one string is present in another character by character using string.contains(). If so, the function returns true or false. It helps find the parameter reduction. It is used if Instr() fails to find the occurrence.

Algorithm 3 Algorithm Reduction_Discernibility_function(Discernibityfn_matrix[],q)

```
1: // q is the total count of distinct entries in the Discernibilityfn_matrix[]
2: for i ← 1 to q do
3:   S = Discernibityfn_matrix[i]
4:   for k ← i + 1 to q do
5:          if (k <> 1) then
6:                 If instr(Discernibityfn_matrix[ik], s) <> 0 or check(Discernibityfn_
                   matrix[k], s) = true then
7:                        Discernibityfn_matrix[k] = “”
8:                 end if
9:          end if
10:         Next k
11: end for
12: Next i
13: end for
14: m = 1
15: for i ← 1 to q do
16: if Discernibityfn_matrix[i] <> “” then
17:         Reduction[m] = Discernibityfn_matrix[i]
18:         m = m + 1
19: end if
20: Next i
21: end for
22: for i ← 1 to m do
23: Display Reduction[i]
24: Next i
25: end for
```

Algorithm 4 Algorithm Check(S1, S2)

```
1: // S1 and S2 are two string entries form Discernibilityfn_matrix[][]
2: //Algorithm returns Yes if either S1 is in S2 or S2 is in S1
3: i = s1.length
4: j = s2.length
5: c1[] = s1.tochararray()
6: c2[] = s2.tochararray()
7: if i > j) then
8:    min = i
9: else
10: min = j
11: end if
12: if i <= j then
13:  for k ← 0 to i − 1 do
14:          if s2.Contains(c1(k)) then
15:                  count = count + 1
16:          end if
17:          Next k
18: end for
19: else
20: for k ← 0 To j − 1 do
21:          if s1.Contains(c1(k)) then
22:                  count = count + 1
23:          end if
24:          Next k
25: end for
26: end if
27: if count = min then
28: return Yes
29: else
30: return No
31: end if
```

The running time performance of the algorithm Reduction_Discernibility_function is $O(q^2)$, where q defines the order of the disceribilityfn_matrix. Also, it includes the time taken to complete the execution of the function call. The major part of the check function is the execution of contains(), which is repeated for the times, and it is equivalent to the total count of the characters in the string that is minimum among two strings passed. Therefore, $O(q^2)$ gives the complete complexity of the procedure.

RESULTS AND DISCUSSION

The developed procedure is realized to identify the key Features that are to be applied in the upcoming analysis of ailment of the affected from innumerable medicinal descriptions comprising lipid profile.

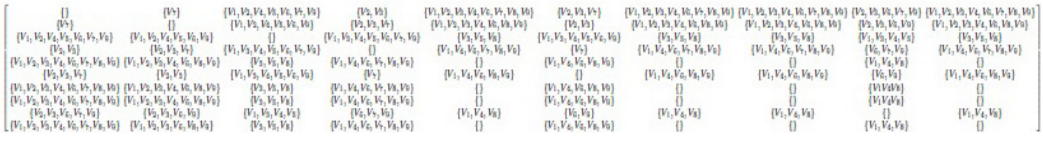
Parameter Reduction on Cardiovascular Disease Patient Data Set

For our ease, we consider P = {E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9, E_{10}} to be the set of patients and V = {V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9} be the different levels of values of the parameters blood glucose (mg/dl), blood urea (mg/dl), serum creatinine (mg/dl), total Cholesterol (mg/dl), HDL cholesterol (mg/dl), LDL Cholesterol (mg/dl), triglycerides (mg/dl), risk factor-1 and risk factor-II respectively that each patient in P has. The standard levels are 70–110 (mg/dl), 10–45 (mg/dl), 0.6–1.5 (mg/dl), < 200 (mg/dl), > 35 (mg/dl), < 130 (mg/dl), < 200 (mg/dl), 3.3–4.9, 1.2–3.5 for blood glucose V1 (mg/dl), blood urea V_2 (mg/dl), serum creatinine V_3 (mg/dl), total cholesterol V_4 (mg/dl), HDL cholesterol V_5 (mg/dl), LDL cholesterol V_6 (mg/dl), triglycerides V_7 (mg/dl), risk factor-I V_8 and Risk factor-II V_9 respectively (Table 1).

Table 1
Several patient's data

Patients / Parameters	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
V_1	198	302	74	256	108	245	98	98	148	70
V_2	65	75	20	45	26	26	26	26	30	22
V_3	2.5	1.6	0.5	1.4	0.7	0.8	0.8	0.9	1.2	0.6
V_4	360	254	175	252	186	235	174	165	202	152
V_5	50	45	33	42	38	42	45	45	40	35
V_6	250	169	110	168	113	153	97	84	124	83
V_7	298	198	156	206	175	198	156	178	186	168
V_8	7.2	5.6	5.3	6	4.9	5.6	3.9	3.7	5.1	4.3
V_9	5	3.8	3.3	4	3	3.6	2.2	1.9	3.1	2.4

It gives the discernibility matrix D =



Then, the discernibility function f_A is given by

$\triangle f_A = \{V_7\} \wedge \{V_1, V_2, V_4, V_5, V_6, V_7, V_9\} \wedge \{V_2, V_3\} \wedge \{V_1, V_2, V_3, V_4, V_6, V_7, V_8, V_9\} \wedge \{V_2, V_3, V_7\} \wedge \{V_2, V_3, V_6, V_7, V_9\} \wedge \{V_1, V_2, V_4, V_5, V_6, V_9\} \wedge \{V_1, V_2, V_3, V_4, V_6, V_8, V_9\} \wedge \{V_2,$

$V_3, V_6, V_9\} \wedge \{V_1, V_3, V_4, V_5, V_6, V_7, V_9\} \wedge \{V_1, V_4, V_6, V_7, V_8, V_9\} \wedge \{V_6, V_7, V_9\} \wedge \{V_3, V_5, V_8\} \wedge \{V_1, V_4, V_6, V_8, V_9\} \wedge \{V_1, V_4, V_8\} \wedge \{V_1, V_3, V_4, V_5, V_6, V_9\} \wedge \{V_6, V_9\} \wedge \{V_1, V_3, V_4, V_5\} \wedge \{V_1, V_2, V_3, V_6, V_8, V_9\}$.

Here, the various parameter reductions returned as output from the proposed soft-set-based parameter reduction algorithms are taken as the data set for which the standard machine learning classifiers are applied, the model accuracies are evaluated, and a comparison is made to identify the predominant parameters with better accuracy.

In a hybrid approach, all 11 independent data set parameters are considered for reductions (Table 2). For the convenience of implementation, the parameters are labeled using the alphabet indicated in Table 2.

The data set is represented as a soft-set, and a sample of the data set is passed as inputs to the proposed soft-set parameter reduction algorithms—the set of parameter reductions returned as output is indicated in Table 3.

Table 2
All 11 parameters with label

Label	Parameters
A	Gender
B	Blood−Hemoglobin
C	Sugar−Fasting
D	Sugar−PP
E	Urea
F	Creatinine
G	Total−Cholesterol
H	Triglycerides
I	HDL−Cholesterol
J	LDL−Cholesterol
K	VLDL−Cholesterol

Table 3
List of parameter reductions

S. No	Parameter Reductions
1.	BCFGIJK
2.	BCFGHIJK
3.	CFGHIJK
4.	ACFGIK
5.	ACFGIJ
6.	ABEGHI
7.	ACEGIJ
8.	ACEGHI
9.	AFGHJK
10.	ACEFGI
11.	BEGHIK
12.	AEGHIJ
13.	AEFGHI
14.	BCEGIK
15.	CEGHIK
16.	ABCEGI
17.	ACEGK
18.	AEGHK

Risk Factor Prediction with Machine Learning Techniques

The cardiovascular disease data set is taken, and a machine learning technique is applied to identify the accuracy of model predictions. In our earlier work, out of 11 independent parameters namely Gender, Blood Hemoglobin, Sugar Fasting, Sugar PP, Urea, Creatinine, Total Cholesterol, Triglycerides, HDL Cholesterol, LDL Cholesterol and V LDL Cholesterol, there occurs positive correlation between Sugar Fasting and Sugar PP with a coefficient equal to 0.847; Creatinine and Urea has correlation coefficient equal to 0.625; LDL cholesterol and Total Cholesterol has

correlation coefficient equal to 0.983; VLDL cholesterol and Triglycerides has correlation coefficient equal to 0.998 (Kannan & Menaga, 2022). Hence, to diagnose further Sugar PP, creatinine, LDL cholesterol, and VLDL cholesterol were not considered. Henceforth, 8 parameters were considered for Risk Forecasting.

Subsequently, various supervised machine learning classification algorithms, namely LDA, KNN, CART, Random Forest, SVM, and Nave Bayes were applied for the data set with the model accuracies as 61.54%, 65.38%, 65.38%, 69.23%, 61.08%, and 57.69% respectively. The result shows that the Random Forest method gives better accuracy than all the applied classifiers.

Now, a hybrid approach is attempted to improve the prediction accuracy. The machine learning classifiers KNN, LDA, SVM, CART, Random Forest, and Naive Bayes are implemented for the above 18 parameter reductions, and the accuracy of model predictions is evaluated.

True positive and negative are dealings of a test's capability that properly categorize an individual with or without illness. True positive mentions a test's capability to designate an individual with illness as positive. A highly sensitive test designates that there are few false negative consequences, and thus, fewer belongings of illness are missed. The true negative test is its ability to designate an individual who does not have an illness. A highly specific test indicates that there are few false positive results. Having specificity as low will not be appropriate for screening since many people without the illness will screen positive and hypothetically collect redundant analytical measures. It is observed that the highest values 0. 8462, 0.9231 are the sensitivity and specificity, respectively, in the parameter reductions ACFGIJ and BCFGIJK, whereas 0.8462 and 0.5485 are the values for sensitivity and specificity in ABCEGHI in Kannan and Menaga (2022). Table 4 recapitulates the true positive and specificity metrics of various classifiers.

Accuracy is one of the measurements for assessing the efficiency of the model's classification. The informal meaning of accuracy is defining the segment of estimates for the right model. Formally, accuracy is defined as $\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$. For two-fold organization, accuracy can also be computed as positives and negatives as follows: $\text{Accuracy} = \frac{\text{T.P+T.N}}{\text{T.P+T.N+F.P+F.N}}$, where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

From Table 5, it is known that improved accuracies are obtained by the Random Forest method. The two-parameter reductions, namely Blood Hemoglobin, Sugar Fasting, Creatinine, Total Cholesterol, HDL Cholesterol, LDL, VLDL (BCFGIJK) and Gender, Sugar Fasting, Creatinine, Total Cholesterol, HDL Cholesterol, LDL (ACFGIJ), provide the highest accuracy 88.46% out of all remaining parameter reductions. However, the

Table 4
Sensitivity and specificity table

Parameters	Sensitivity and specificity metric of various classifiers											
	LDA		KNN		SVM		CART		NB		RF	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
ABCEGHI in (Kannan, 2022)	0.6923	0.5385	0.8462	0.4615	0.7692	0.5385	0.6154	0.6923	1.0000	0.2308	0.8462	0.5385
ACFGIJ	0.5385	0.6923	0.7692	0.6154	0.8462	0.6923	0.6923	0.7692	0.4615	0.7692	0.8462	0.9231
BCFGIJK	0.6923	0.6154	0.6154	0.5385	0.7692	0.7692	0.7692	0.4615	0.5385	0.8462	0.8462	0.9231
ACEGIJ	0.5385	0.4615	0.6923	0.6154	0.9231	0.7692	0.6923	0.7692	0.7692	0.5385	0.7692	0.9231
ABCEGI	0.5385	0.5385	0.6154	0.6923	0.6154	0.6154	0.5385	0.6154	0.6923	0.5385	0.7692	0.8462
BCFGHIJK	0.9231	0.6154	0.7692	0.5385	0.7692	0.5385	0.8462	0.4615	0.7692	0.2308	0.7692	0.7692
AEGHIJ	0.6154	0.6154	0.8462	0.3846	0.9231	0.7692	0.3846	0.4615	0.7692	0.5385	0.6923	0.8462
CFGHIJK	0.4615	0.6923	0.6154	0.5385	0.7692	0.6923	0.6923	0.7692	0.6923	0.5385	0.6923	0.7692
ACEFGI	0.4615	0.6154	0.6154	0.6923	0.6923	0.6154	0.76923	0.07692	0.15385	0.92308	0.6154	0.7692
BEGHIK	0.5385	0.6154	0.6923	0.3077	0.6923	0.3846	0.3846	0.3077	0.5385	0.3077	0.6154	0.6923
ACFGJK	0.5385	0.6923	0.6923	0.6154	0.6923	0.4615	0.6923	0.7692	0.6923	0.6154	0.6154	0.6154
ABEGHI	0.6154	0.6923	0.6923	0.3077	0.8462	0.5385	0.3846	0.3077	0.9231	0.5385	0.6154	0.6154
AEFGHI	0.6154	0.6923	0.6923	0.3077	0.9231	0.5385	0.3846	0.3077	0.5385	0.3077	0.6154	0.6154
ACEGHI	0.6154	0.6154	0.7692	0.5385	0.6923	0.4615	0.6923	0.5385	0.6154	0.3846	0.5385	0.6154
AFGHJK	0.4615	0.6923	0.7692	0.3846	0.7692	0.3846	0.3846	0.4615	0.6154	0.5385	0.5385	0.6154
BCEGIK	0.5385	0.5385	0.6154	0.5385	0.6923	0.5385	0.5385	0.6154	0.5385	0.3846	0.6154	0.5385
CEGHIK	0.6154	0.6154	0.7692	0.5385	0.7692	0.5385	0.6923	0.5385	0.6154	0.3846	0.5385	0.6154
ACEGK	0.5385	0.4615	0.6923	0.5385	0.5385	0.4615	0.6923	0.5385	0.6923	0.3846	0.4615	0.4615
AEGHK	0.7692	0.4615	0.6923	0.3077	0.6923	0.3846	0.3846	0.3077	0.5385	0.3846	0.3846	0.4615

Table 5
Accuracy

Accuracy	Metric of various classifiers					
Parameters	LDA in %	KNN in %	SVM in %	CART in %	NB in %	RF in %
ABCEGHI in (Kannan, 2022)	61.54	65.38	65.38	65.38	61.34	69.23
ACFGIJ	61.54	69.23	76.92	73.08	61.54	88.46
BCFGIJK	61.16	57.69	76.92	61.54	69.23	88.46
ACEGIJ	50	63.58	84.62	73.08	63.58	84.62
ABCEGI	53.85	63.58	61.54	57.69	61.54	80.77
BCFGHIJK	76.92	61.54	73.08	53.08	69.23	76.92
AEGHIJ	61.54	61.54	84.62	42.31	65.38	76.92
CFGHIJK	57.69	57.69	73.08	73.08	61.54	73.08
ACEFGI	53.85	65.38	65.38	42.31	53.85	69.23
BEGHIK	57.69	50	53.85	34.62	42.31	65.38
ACFGJK	61.54	65.38	57.69	73.08	65.38	61.54
ABEGHI	65.38	50	69.23	34.62	73.08	61.54
AEFGHI	65.38	50	73.08	34.62	42.31	61.54
ACEGHI	61.54	65.38	57.69	61.54	50	57.69
AFGHJK	57.69	57.69	57.69	42.31	57.69	57.69
BCEGIK	53.85	57.69	61.54	57.69	46.15	57.69
CEGHIK	61.54	65.38	65.38	61.54	50	57.69
ACEGK	50	61.54	50	61.54	53.85	46.15
AEGHK	61.54	50	53.85	34.62	46.15	42.31

accuracies of LDA, KNN, SVM, CART, and RF in ACFGIJ are improved than that of LDA, KNN, SVM, CART, NB, and RF in BCFGIJK respectively, whereas in NB, the accuracy is decreased. At the same time, the accuracies of LDA, KNN, SVM, CART, NB, and RF in ACFGIJ are either improved or remain the same as that of LDA, KNN, SVM, CART, NB, and RF in ABCEGHI in Kannan and Menaga (2022), respectively. Hence, the proposed hybrid approach drastically improves accuracy over the conventional ACFGIJ and BCFGIJK methods (Figures 2, 3, and 4).

Moreover, Sugar Fasting (C), Total Cholesterol (G), and HDL Cholesterol (I) are the core components identified in all these three parameter reductions. Then, the importance of the parameters is identified (Random Forest) in BCFGIJK and ACFGIJ, shown in Tables 6 and 7, respectively.

Initially, the order of importance of the parameters in BCFGIJK [Figure 5(a)] is HDL cholesterol, LDL cholesterol, Total cholesterol, VLDL cholesterol, Creatinine, Blood Hemoglobin, and Sugar fasting, whereas, after tuning, it is LDL cholesterol, HDL

cholesterol, Total cholesterol, VLDL cholesterol, Creatinine, Blood Hemoglobin and Sugar fasting. Similarly, the order of importance of the parameters in ACFGIJ (Figure 5b) is HDL cholesterol, LDL cholesterol, Total cholesterol, Creatinine, Gender, and Sugar fasting, whereas, after tuning, it is LDL cholesterol, HDL cholesterol, Total cholesterol, Sugar fasting, Creatinine and Gender.

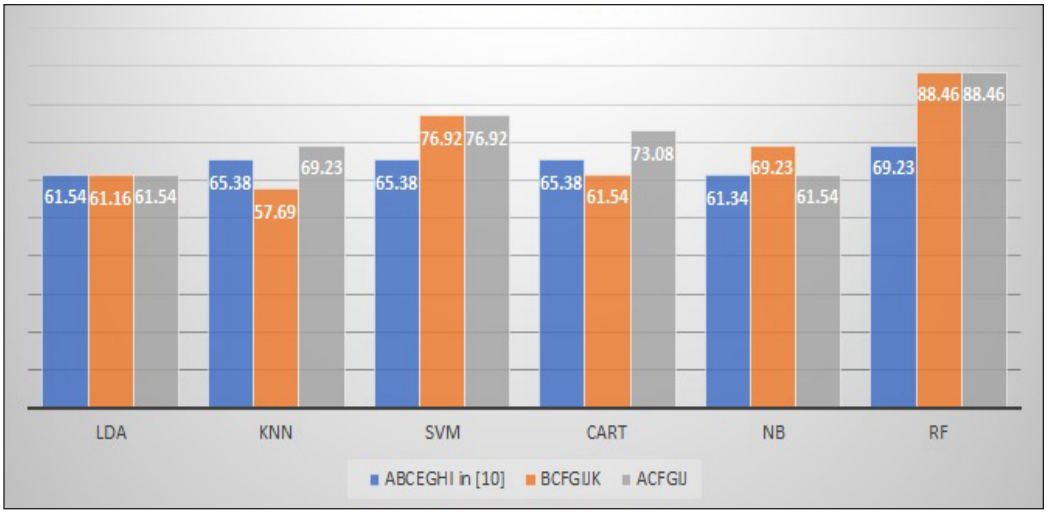
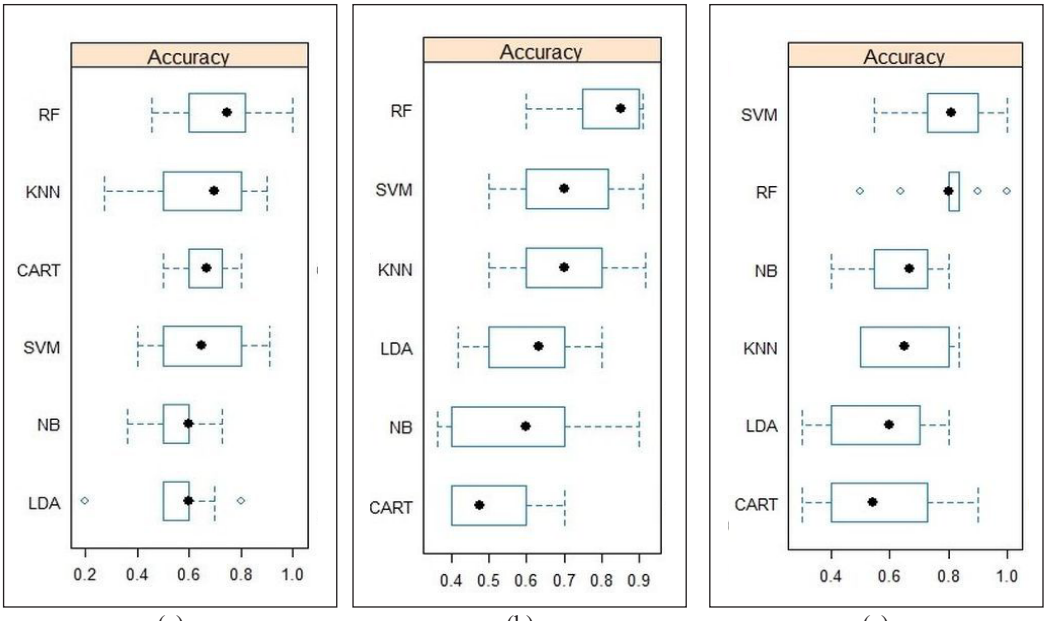
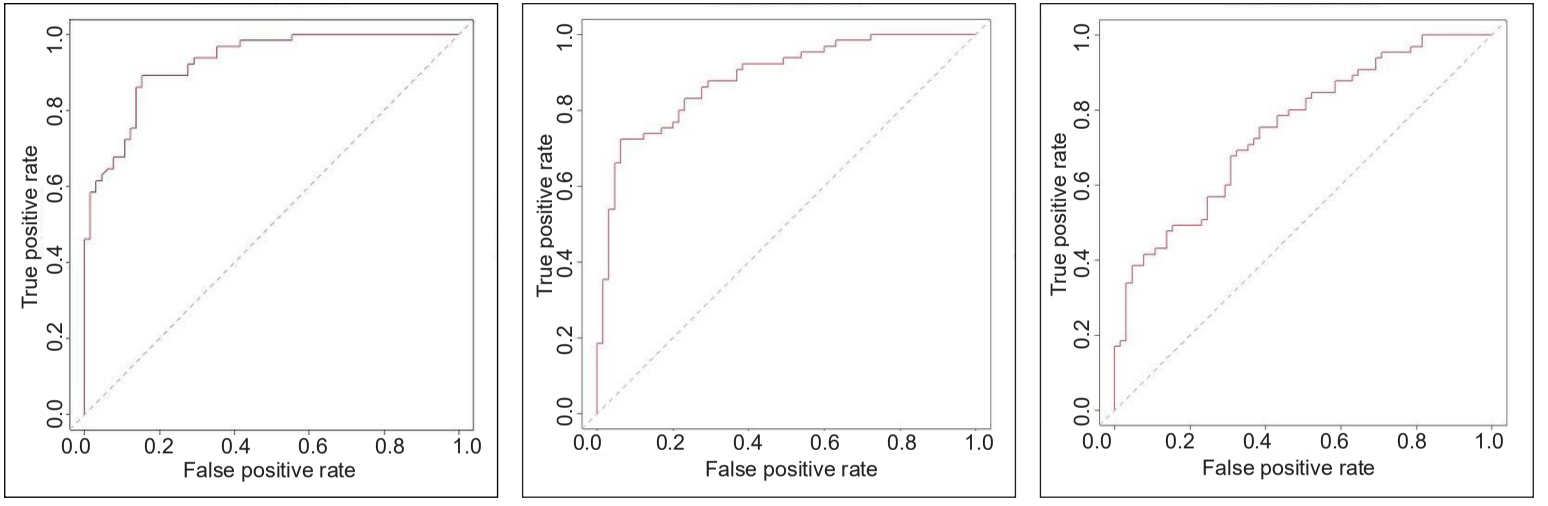


Figure 2. Comparison of accuracy



(a) (b) (c)

Figure 3. Comparison of accuracy: (a) ABCEGHI in (Kannan, 2022); (b) BCFGIJK; and (c) ACFGIJ



(a) (b) (c)

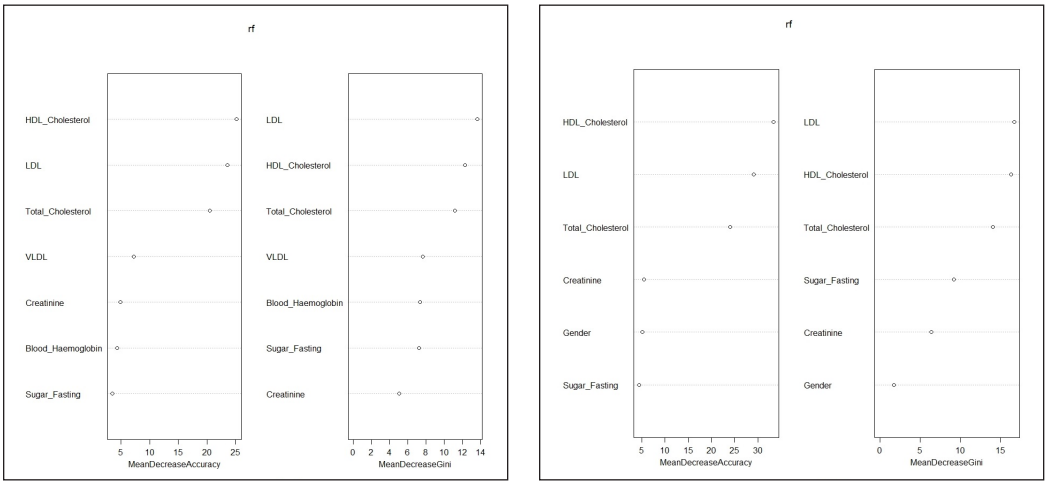
Figure 4. ROC Curves: (a) ABCEGHI in (Kannan, 2022); (b) BCFGIJK; and (c) ACFGIJ

Table 6
Importance of parameters in random forest in BCFGIJK

Parameters	There is no risk	There is risk	Mean decrease accuracy	Mean decrease Gini
Blood Hemoglobin	4.7493817	1.337873	4.352876	7.332752
Sugar Fasting	0.8752779	4.387068	3.477273	7.237084
Creatinine	4.0794305	2.948084	4.899098	5.083962
Total Cholesterol	15.8845728	16.667775	20.519157	11.189501
H.D.L Cholesterol	22.1786327	16.667775	25.162436	12.299851
L.D.L Cholesterol	17.9061889	17.977023	23.582813	13.653628
V.L.D.L Cholesterol	1.6163369	7.961650	7.252641	7.665264

Table 7
Importance of parameters in random forest in ACFGIJ

Parameters	There is no risk	There is risk	Mean decrease accuracy	Mean decrease Gini
Gender	0.7070882	6.234695	5.186608	1.745934
Sugar Fasting	1.6658202	4.722036	4.421667	9.171790
Creatinine	3.1746004	5.079762	5.521531	6.397909
Total Cholesterol	17.3298524	17.734249	24.069675	14.070509
H.D.L Cholesterol	26.4972019	26.996690	33.388424	16.315529
L.D.L Cholesterol	19.8903453	23.907039	29.075341	16.724294



(a) (b)

Figure 5. Importance of parameters in: (a) BCFGIJK; and (b) ACFGIJ

CONCLUSION

The present paper proposes an algorithm to minimize the parameters based on a soft-set discernibility matrix. The execution time analysis of the suggested procedure is calculated. Further, parameter reduction algorithms are implemented in a real-time decision-making application, i.e., for predicting cardiovascular disease risk factors. Also, a hybrid approach combining soft-set and machine learning techniques is proposed for efficient parameter reduction, and accuracy comparisons were made. In the existing approach, the Random Forest provided better accuracy (69.23%), and the proposed hybrid approach provided improved accuracy (88.46%) for the same classifier, which is a drastic improvement. A more comparative analysis between multiple datasets and their properties might be undertaken to find all the essential elements for forecasting the risk factor. Furthermore, rather than conventional machine learning approaches, deep learning techniques like artificial neural networks and structural equation modeling can be used in the future.

ACKNOWLEDGEMENT

The authors thank Dr. D. Rajalakshmi, Department of Mathematics, Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam, Tamil Nadu, INDIA, for providing the data set to enhance the quality of the paper.

REFERENCES

Abirami, R. S., & Kumar, G. S. (2022). Comparative study based on analysis of coronavirus disease (COVID-19) detection and prediction using machine learning models. *SN Computer Science, 3*(1), Article 79. https://doi.org/10.1007/s42979-021-00965-2

AlArjani, A., Nasseef, M. T., Kamal, S. M., Rao, B. V. S., Mahmud, M., & Uddin, M. S. (2022). Application of mathematical modeling in prediction of COVID-19 transmission dynamics. *Arabian Journal for Science and Engineering, 47*(8), 10163-10186. https://doi.org/10.1007/s13369-021-06419-4

Ali, M. I. (2012). Another view on reduction of parameters in soft sets. *Applied Soft Computing, 12*(6), 1814-1821. https://doi.org/10.1016/j.asoc.2012.01.002

Alotaibi, A. S. (2022). Hybrid model based on relief algorithm and k-nearest neighbor for erythemato-squamous diseases forecasting. *Arabian Journal for Science and Engineering, 47*(2), 1299-1307. URL https://doi.org/10.1007/s13369-021-05921-z

Ansari, G. A., Bhat, S. S., Ansari, M. D., Ahmad, S., Nazeer, J., & Eljialy, A. E. M. (2023). Performance evaluation of machine learning techniques (MLT) for heart disease prediction. *Computational and Mathematical Methods in Medicine, 2023*, Article 8191261. https://doi.org/10.1155/2023/8191261

Bhat, S. S., Selvam, V., Ansari, G. A., Ansari, M. D., & Rahman, M. H. (2022). Prevalence and early prediction of diabetes using machine learning in North Kashmir: A case study of District Bandipora. *Computational Intelligence and Neuroscience, 2022*, Article 2789760. https://doi.org/10.1155/2022/2789760

Boukhatem, C., Youssef, H. Y., & Nassif, A. B. (2022). Heart disease prediction using machine learning. In *2022 Advances in Science and Engineering Technology International Conferences (ASET)* pp. 1-6. IEEE Publishing. https://doi.org/10.1109/ASET53988.2022.9734880

Cagman, N., & Enginoglu, S. (2010). Soft set theory and uni-int decision making. *European Journal of Operational Research, 207*(2), 848-855. https://doi.org/10.1016/j.ejor.2010.05.004

Chen, D., Tsang, E. C. C., Yeung, D. S., & Wang, X. (2005). The parameterization reduction of soft sets and its applications. *Computers & Mathematics with Applications, 49*(5), 757-763. https://doi.org/10.1016/j.camwa.2004.10.036

Gupta, G., Gupta, A., Jaiswal, V., & Ansari, M. D. (2018). A review and analysis of mobile health applications for alzheimer patients and caregivers. In *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 171-175). IEEE Publishing. https://doi.org/10.1109/PDGC.2018.8745995

Herawan, T., Rose, A. N. M., & Deris, M. M. (2009). Soft set theoretic approach for dimensionality reduction. In D. Ślęzak, T. Kim, Y. Zhang, J. Ma & K. Chung (Eds.), *Database Theory and Application* (pp. 171-178). Springer. https://doi.org/10.1007/978-3-642-10583-8_20

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing. https://doi.org/10.1088/1757-899X/1022/1/012072

Kannan, K., & Menaga, A. (2022). Some efficient algorithms on the parameter reduction of soft sets for decision making problems. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 92*(3), 319-329. https://doi.org/10.1007/s40010-021-00730-3

Khan, A., & Zhu, Y. (2019). An improved algorithm for normal parameter reduction of soft set. *Journal of Intelligence and Fuzzy Systems, 37*(2), 2953-2968. https://doi.org/10.3233/JIFS-190071

Kong, Z., Gao, L., Wang, L., & Li, S. (2008). The normal parameter reduction of soft sets and its algorithm. *Computers and Mathematics with Applications, 56*(12), 3029-3037. https://doi.org/10.1016/j.camwa.2008.07.013

Kong, Z., Zhang, G. D., & Wang, L. F. (2014). Normal parameter reduction in soft set based on harmony search algorithm. *Applied Mechanics and Materials, 644-650*, 2173-2176. https://doi.org/10.4028/www.scientific.net/AMM.644-650.2173

Kovkov, D. V., Kolbanov, V. M., & Molodtsov, D. A. (2007). Soft sets theory-based optimization. *Journal of Computer and Systems Sciences International, 46*(6), 872-880. https://doi.org/10.1134/S1064230707060032

Ma, X., Sulaiman, N., Qin, H., Herawan, T., & Zain, J. M. (2011). A new efficient normal parameter reduction algorithm of soft sets. *Computer & Mathematics with Applications, 62*(2), 588-598. https://doi.org/10.1016/j.camwa.2011.05.038

Maji, P. K., Roy, A. R., & Biswas, R. (2002). An application of soft sets in a decision making problem. *Computers & Mathematics with Applications, 44*(8), 1077-1083.

Mezzatesta, S., Torino, C., Meo, P. D., Fiumara, G., & Vilasi, A. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. *Computer Methods and Programs in Biomedicine, 177*, 9-15. https://doi.org/10.1016/j.cmpb.2019.05.005

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mohammed, M. A. T., Sadiq, A. S., Arshah, R. A., Ernawan, F., & Mirjalili, S. M. (2017). Soft set decision/forecasting system based on hybrid parameter reduction algorithm. *Journal of Telecommunication, Electronic and Computer Engineering, 9*(2-7), 143-148.

Molodtsov, D. (1999). Soft set theory - First results. *Computers & Mathematics with Applications, 37*(4-5), 19-31. https://doi.org/10.1016/S0898-1221(99)00056-5

Muhammad, L. J., Al-Shourbaji, I., Haruna, A. A., Mohammed, I. A., Ahmad, A., & Jibrin, M. B. (2021). Machine learning predictive models for coronary artery disease. *SN Computer Science, 2*(5), Article 350. https://doi.org/10.1007/s42979-021-00731-4

Palaniappan, S., Raghavi, V., David, B., & Nisha, S. P. (2021). Prediction of epidemic disease dynamics on the infection risk using machine learning algorithms. *SN Computer Science, 3*(1), Article 47. https://doi.org/10.1007/s42979-021-00902-3

Pawlak, Z. (1982). Rough sets. *International Journal of Computer & Information Sciences, 11*(5), 341-356. https://doi.org/10.1007/BF01001956

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data (Vol. 9)*. Springer Science & Business Media.

Rahman, M. M., Islam, M. M., Manik, M. M. H., Islam, M. R., & Al-Rakhami, M. S. (2021). Machine learning approaches for tackling novel Coronavirus (COVID-19) pandemic. *SN Computer Science, 2*(5), Article 384. https://doi.org/10.1007/s42979-021-00774-7

Rose, A. N. M., Herawan, T., & Deris, M. M. (2010). A framework of decision making based on maximal supported sets. In L. Zhang, B. L. Lu & J. Kwok (Eds.), *Advances in Neural Networks - ISNN 2010* (pp. 473-482). Springer. https://doi.org/10.1007/978-3-642-13278-0_61

Sadiq, A. S., Tahir, M. A., Ahmed, A. A., & Alghushami, A. (2020). Normal parameter reduction algorithm in soft set based on hybrid binary particle swarm and biogeography optimizer. *Neural Computing and Applications, 32*(16), 12221-12239. https://doi.org/10.1007/s00521-019-04423-2

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science, 1*(6), Article 345. https://doi.org/10.1007/s42979-020-00365-y

Sharma, D., Gotlieb, N., Farkouh, M. E., Patel, K., Xu, W., & Bhat, M. (2022). Machine learning approach to classify cardiovascular disease in patients with nonalcoholic fatty liver disease in the UK Biobank Cohort. *Journal of the American Heart Association, 11*(1), Article e022576. https://doi.org/10.1161/JAHA.121.022576

Wang, S. H., Wu, K., Chu, T., Fernandes, S. L., Zhou, Q., Zhang, Y. D., & Sun, J. (2021). SOSPCNN: Structurally optimized stochastic pooling convolutional neural network for tetralogy of fallot recognition. *Wireless Communications and Mobile Computing, 2021*, Article 5792975. https://doi.org/10.1155/2021/5792975

Xie, N. X. (2016). An algorithm on the parameter reduction of soft sets. *Fuzzy Information and Engineering, 8*(2), 127-145. https://doi.org/10.1016/j.fiae.2016.06.001

Yuksel, S., Dizman, T., Yildizdan, G., & Sert, U. (2013). Application of soft sets to diagnose the prostate cancer risk. *Journal of Inequalities and Applications, 2013*(1), Article 229. https://doi.org/10.1186/1029-242X-2013-229

Zhang, Y., Lu, S., Zhou, X., Yang, M., Wu, L., Liu, B., Phillips, P., & Wang, S. (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: Decision tree, k-nearest neighbors, and support vector machine. *Simulation, 92*(9), 861-871. https://doi.org/10.1177/0037549716666962