Original Article

An Improved Squirrel Search Optimization for Medical Data Classification Model

[1]M. RAJA [2]M. Y. MOHAMED PARVEES,
[1]Research Scholar, Department of Computer and Information science, Annamalai University, Annamalainagar – 608002, India. mraja.cdm@gmail.com
[2]Department of Computer and Information science, Annamalai University, Annamalainagar – 608002, India. yparvees@gmail.com
DOI: 10.47750/pnr.2022.13.S06.255

Abstract

At present times, computer aided diagnosis (CAD) models become familiar in the healthcare sector. Medical diagnosis is identified to be subjective and it is based on the available data as well as physician's experience. Machine learning (ML) techniques are commonly employed to design effective CAD models due to its stronger ability of identifying the complicated relationship in the medical data. This paper aims to develop a novel data classification model using squirrel search algorithm (SSA) with Mode Ranking method called MRISSA for healthcare diagnosis. Experiments were done on heart disease datasets available in Kaggle to evaluate the performance of the suggested technique. The outcomes show how effective the hybrid (MRISSA+ XGBoost) strategy is improved.

Keywords: Medical data classification, Machine learning, Mode Ranking

1. Introduction

Basically, medical diagnosis tends to enhance the human lifespan which depends upon the clinical and non-clinical outline. The earlier disease diagnosis is more significant and essential in healthcare application where the abnormality is examined and treated accordingly [1]. However, providing a complete medical report is impossible for medical experts which are performed using numerous clinical tools [2]. Followed by, smart clinical models are composed of voluminous details where clinical diagnosis have partial data, instability, and unreliable information, that is required for resolving the issues involved in medical diagnosis [3]. Due to the disease severity and inexistence of medical experience regarding the problems, accurate data is not available as clinical analyzing process is comprised of irregularities, and inefficiency. Thus, uncertainty is one of the basic factors which affect the quality of clinical data [4, 5].

The clinical information showcases exclusive features along with the noise that results in manual and logical failures, missing values as well as sparseness. The supremacy of a data shows better impact in the superiority of mining results [6]. In recent times, severe neuro cognitive infections are examined as a serious disorder. Accurate and well-trained analysis of these infections are important as they are linked to diverse results with enhanced management of under-recognized neuropsychiatric presentations, which are assumed as a major healthcare target [7]. In order to resolve these issues, diverse corrective solutions have been employed. An effective disease analyzing module is named as sequential diagnosis is developed, which contributes in resolving the multifaceted decision crisis [8]. Similarly, supervised learning models has been used to map the input data to output (labels) from a collection of training instances, that has ensured a better efficiency in clinical analysis. Supervised quantification methods are used in healthcare analysis, and disease prediction operations [9]. Furthermore, medical support system or clinical decision support system has been developed in medical sciences for assisting physicians in making clinical decisions, especially in finding the type of disease by examining the symptoms effectively [10].

Studies in computerized intelligent systems for clinical domains are essential. Actually, medical experts collect the patient details according to the patients' signs and name the disease. Besides, prognostic relevance of symptoms in specific diseases and diagnostic precision of a patient depends upon the knowledge of medical experts. Nowadays, medical experience and services are developed in a rapid manner, for instance, new medicines and drugs were introduced and regular maintenance of patient's details with accuracy is highly

complex for the medical experts [11]. Followed by, using the advent of computing methods, it is simple to gather the data in digital fashion, for sample, in dedicated databases of electronic patient details. Next, the developments of programmatic clinical decision support systems are one of the feasible models to guide the doctors for accurate disease diagnosis [12]. Regardless, massive problems are solved before developing an effective clinical decision support system where an appropriate solution can be deployed under the existence of uncertainty as well as inaccuracy [13].

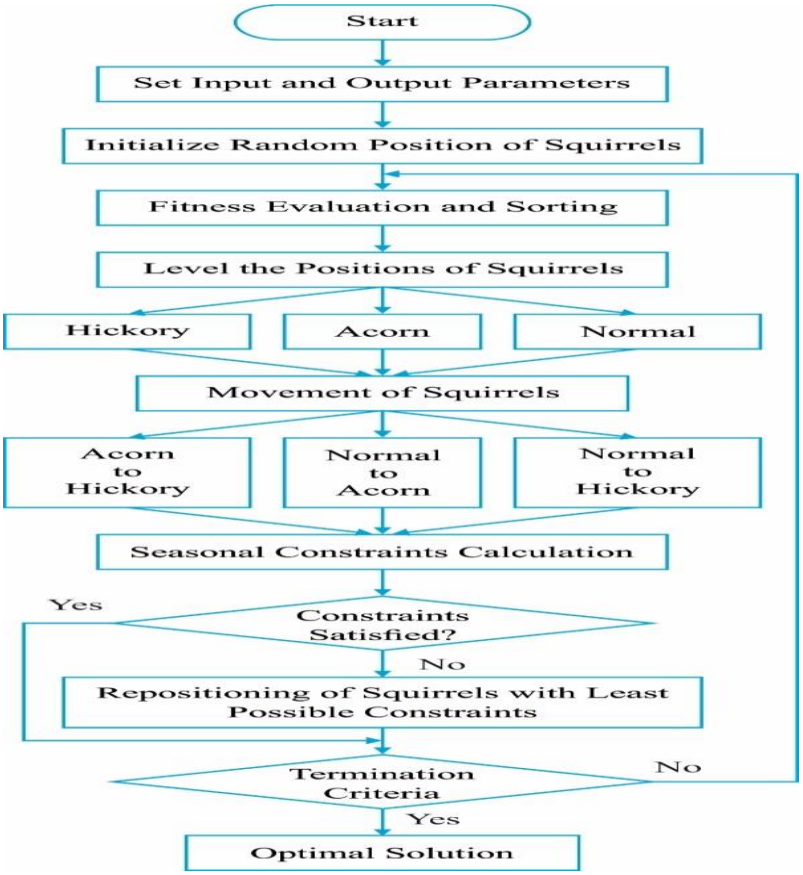
As clinical experts' experience and knowledge are highly essential, examining the patient's condition is carried out using a proficient model named as Machine Learning (ML) [14]. Here, the clinical experts have applied computerized intelligent systems to decision support such as surgical imagery and X-ray photography. When the patient undergoes a treatment, a doctor has to find the major cause of a disease under the examination of symptoms and

assures the disease and offer the treatment accordingly [15]. Concurrently, computerized intelligent systems are applicable for physician to make a robust and effective decision, for sample, by learning from previous cases in a large-scale database of electronic patient records with appropriate justifications. The benefits of applying intelligent systems ate enhanced accuracy in s disease analysis, and simultaneously, reduced time and costs related with patient recovery [16].

The ML approaches were deployed to support diverse clinical decision making operations. Then, intelligent classification models were employed to prognosis, diagnosis, diabetes screening, breast cancer prediction and Parkinsons disease [17]. The count of neural-fuzzy schemes are employed as classifiers in heart disease, as it is suitable for learning the data samples and generalize the training samples [18]. Some of the classifiers are Fuzzy Neural Networks (FNN), fuzzy probabilistic neural networks, and fuzzy learning vector quantization systems [19]. Hence, a disadvantage of this approach is that it is not capable to define the predictions [20]. Finally, the central premise of this study is to make a ML-relied system as potential one to resolve the input case, and to offer a possible definition for disease prediction.

2. The Proposed ISSA-MR Model

The working process involved in the proposed ISSA-MR model. As depicted, the input medical data is initially preprocessed to enhance the data quality. Next, ISSA-MR based classification model is applied to identify the presence of disease. The detailed workingprocesses of these models are discussed in the subsequent sections.



Block diagram of ISSA-MR Model

2.1. Preprocessing

Initially, the input medical data in any .arff format is converted into CSV format. Then, data normalization process is executed using minimum-maximum (min-max) technique. At this point, the maximum and minimum values in the data gathering are assumed. Each data is normalized to the interval of [0, 1]. The aim is to fix the least value to 0 and highest value to

1. It is utilized for defining the procedure involved in min.-max normalization. At last, the class labeling procedure is carried out where the medical data instances are assigned to appropriate classes.

2.2 Basic Squirrel Search Algorithm

SSA mimics the dynamic foraging behavior of southern flying squirrels via gliding, an effective mechanism used by small mammals for travelling long distance, in deciduous forest of Europe and Asia . During warm weather, the squirrels change their locations by gliding from one tree to another in the forest and explore for food resources. They can easily find acorn nuts for meeting daily energy needs. After that, they begin searching hickory nuts (the optimal food source) that are stored for winter. During cold weather, they become less active and maintain their energy requirements with the storage of hickory nuts. When the weather gets warm, flying squirrels become active again. The abovementioned process is repeated and continues throughout the life space of the squirrels, which serves as a foundation of the SSA. According to the food foraging strategy of flying squirrels, the optimization SSA can be modeled by the following phases mathematically.

2.3. Initialize the Algorithm Parameters

The main parameters of the SSA are the maximum number of iteration , the population size , the number of decision variables *n*, the predator presence probability , the scaling factor , the gliding constant , and the upper and lower bounds for decision variable and . These parameters are set in the beginning of the SSA procedure.

2.4. Initialize Flying Squirrels' Locations and Their Sorting

The flying squirrels' locations are randomly initialized in the search apace as follows: where is a uniformly distributed random number in the range .The fitness value of an individual flying squirrel's location is calculated by substituting the value of decision variables into a fitness function: Then the quality of food sources defined by the fitness value of the flying squirrels' locations is sorted in an ascending order: After sorting the food sources of each flying squirrel's location, three types of trees are categorized: hickory tree (hickory nuts food source), oak tree (acorn nuts food source), and normal tree. The location of the best food source (i.e., minimal fitness value) is regarded as the hickory nut tree (), the locations of the following three food sources are supposed to be the acorn nuts trees (), and the rest are considered as normal trees ():

2.5. Generate New Locations through Gliding

Three scenarios may appear after the dynamic gliding process of flying squirrels.

Scenario 1. Flying squirrels on acorn nut trees tend to move towards hickory nut tree. The new locations can be generated as follows where is random gliding distance, is a function which returns a value from the uniform distribution on the interval , and is a gliding constant.

Scenario 2. Some squirrels which are on normal trees may move towards acorn nut tree to fulfill their daily energy needs. The new locations can be generated as follows: where is a function which returns a value from the uniform distribution on the interval .

Scenario 3. Some flying squirrels on normal trees may move towards hickory nut tree if they have already fulfilled their daily energy requirements. In this scenario, the new location of squirrels can be generated as follows:where is a function which returns a value from the uniform distribution on the interval .

In all scenarios, gliding distance is considered to be in the interval between 9 and 20 m . However, this value is quite large and may introduce large perturbations in and hence may cause unsatisfactory performance of the algorithm. In order to achieve acceptable performance of the algorithm, a scaling factor () is introduced as a divisor of and its value is chosen to be 18

2.6. Check Seasonal Monitoring Condition

The foraging behavior of flying squirrels is significantly affected by season variations .Therefore, a seasonal monitoring condition is introduced in the algorithm to prevent the algorithm from being trapped in local optimal solutions.

A seasonal constant and its minimum value are calculated firstly:Then the seasonal monitoring condition is checked. Under the condition of , the winter is over, and the flying squirrels which lose their abilities to explore the forest will randomly relocate their searching positions for food source again:where distribution is a powerful mathematical tool to enhance the global exploration capability of most optimization algorithms [44]:where and are two functions which return a value from the uniform distribution on the interval , is a constant (= 1.5 in this paper), and is calculated as follows:where .

2.7. Stopping Criterion

The algorithm terminates if the maximum number of iterations is satisfied. Otherwise, the behaviors of generating new locations and checking seasonal monitoring condition are repeated.

2.8 Create new positions

By considering the above mentioned facts, 3 scenarios might be feasible in a dynamic scavenging of SQ. In every scenario, it is better to consider that when there is no predator, SQs glide and search effectively in the entire forest and accomplish the favourable food source while if the predator exists, SQs are forced to travel in random walks and identify the adjacent hidden place. The numerical modeling of dynamic foraging embeds the representation as shown in the following.

Scenario 1. The SQs in acorn nut trees (SQ_{oCtl}) walk to the hickory nut tree. Hence, current place of the SQs are exhibited as:

$$SQ^{ni+1} = \begin{cases} SQ^{ni} + di_{gl} \times GL_{cn} \times (SQ^{ni}_{acn} - SQ^{ni}_{hck}) RN_1 & \geq PD_{prb} \\ SQ^{ni}_{nrm} & Arbitrary position Otherwise \end{cases} \quad (10)$$

where di_{gl} implies the gliding distance, RN_1 defines the random value from [0,1], SQ_{hck} refers the location of SQ has reached at hickory nut tree, and ni represents the current round.
GL_{cn} guides in reaching the management among exploration and exploitation. In this method,the measure of GL_{cn} is considered as 1.9.

Scenario 2. In SQs residing in normal trees (SQ_{nrm})walks to the acorn nut trees forsatisfying the regular power demands. Hence, the current location of SQs is referred as:

$$SQ^{ni+1} = \begin{cases} SQ^{ni}_{nrm} + di_{gl} \times GL_{cn} \times (SQ^{ni}_{acn} - SQ^{ni}_{nrm}) RN_2 & \geq PD_{prb} \\ SQ^{ni}_{nrm} & Arbitrary position Otherwise \end{cases} \quad (11)$$

where RN_2denotes the ransom value from [0,1].

Scenario 3. Few SQs on normal trees, and already eaten acorn nuts are developed towards the hickory nut tree for acquiring the hickory nuts that is applied under the inexistence of energy. Here, the current location of SQs is depicted as:

$$SQ^{ni+1} = \begin{cases} SQ^{ni}_{nrm} + di_{gl} \times GL_{cn} \times (SQ^{ni}_{hck} - SQ^{ni}_{nrm}) RN_3 & \geq PD_{prb} \\ SQ^{ni}_{nrm} & Arbitrary position Otherwise \end{cases} \quad (12)$$

where RN_3signifies the random value from [0,1].

2.9 Aerodynamics of gliding

The gliding principles of SQs are implied with respect to equilibrium glide (EG). In EG, inclusion of the lift (L_f) drag (D_r) force develops a resultant force (R_f). The R_f is a magnitude is which equivalent and opposite direction to weight of a SQ(VV_{SQ}). Hence, the SQs provide a linear gliding path with constant velocity (CV) . A SQ gliding at constant speed goes down at an angle θ horizontally. The L_f to D_r proportion is defined as given in (13).

$$\frac{L_f}{D_r} = \frac{1}{\tan\theta} \quad (13)$$

The SQs make an enhanced glide-path length by developing tiny glide angle (θ). The improved glide-path length maximizes the ratio of L_f to D_r. A L_f results in downward deflection of air whereas the passage over wings are determined as given in (14).

$$L_f = 1/2\rho C_{L_f} (CV)^2 S_A \quad (14)$$

Where ρrefers the density of air $(1.204 kg/m^3)$, C_L demonstrates the lift coefficient, CV implies speed $(5.25 m/s)$, and S_A denotes s the surface area of body. The frictional $D_r$is represented as per (15).

$$D_r = 1/2(CV)^2 S_A C_{FD_r} \quad (15)$$

where C_{FD_r} showcases the frictional drag coefficient. In lower speeds, the drag component is dominant always. In case of higher speeds, dominance of component becomes lower. Therefore, from (13), the θ at steady-state as defined below.

$$\theta = \arctan\left(\frac{D_r}{L_f}\right) \quad (16)$$

The randomly evaluated $di_{gl}$is determined as given in (17).

$$d_{gl} = (h_{gl}/\tan\theta) \quad (17)$$

where h_{gl}represents the loss in height behind the gliding [22]. The SQ is composed of ability in adjusting the glide length or d_{gl} by changing the $L_f toD_r$ ratio and considers the essential landing position. The value of d_{gl} is scaled down to reach suitable results. Hence, scaling down embeds the division by a non-zero numeral named as scaling factor (sc_f). Differences in C_{L_f}among 0.675 and 1.5 is considerable while C_{D_r} is equalized as 0.60.

2.10Seasonal monitoring condition

Occasional variants influence the foraging nature of SQs effectively. Therefore, actions of SQs ate impacted by diverse environmental conditions. By considering the behaviour might be realistic towards optimization. Thus, addition of seasonal supervising condition in a method secures the model from trapping in local optima(s).

Estimation of seasonal constant (Se_c)

$$S = \sqrt{\sum_{m=1}^{c} \left(SQ_{acn,m}^{ni} - SQ_{hck,m}^{ni}\right)^2} for \forall tm = 1,2,\ldots \quad (18)$$

Verify of the seasonal monitoring criteria

It should be assured that

$$Se_c^{tm} \prec Se_{min} \quad (19)$$

where $Se_{\min}$ denotes the lower feasible value of seasonal constant. It is determined as:

$$Se_{\min} = \frac{10E^{-6}}{(365)^{ni/(ni_{\max}/2.5)}} \quad (20)$$

where ni and $ni_{\max}$ refers the on-going and higher iteration value, correspondingly.

2.10.1 Arbitrary relocation after completion of the winter season

Here, seasonal monitoring condition is meant to be true (after winter season), the SQs are allocated randomly which is not applicable to identify the food source in winter season. The relocation of SQs is carried out as per (21).

$$SQ^{new}_{nrm} = SQ_{l} + Lé(t) \times (SQ_{w} - SQ_{l}) \tag{21}$$

where Lévy distribution enforces the effective exploration of search space. This distributionis represented as given in (22).

$$(s) \sim |s|^{-1-\alpha} \tag{22}$$

Where $0 \prec \alpha \leq 2$implies an index. Arithmetical function of the distribution is (23).

$$L(s,\beta,\mu) = \begin{cases} \sqrt{\frac{\beta}{2\pi}} \exp\left(\frac{-\beta}{2(s-\mu)}\right) \frac{1}{(s-\mu)^{1.5}} & 0 \prec \mu \prec s \prec \infty \\ 0 & otherwise \end{cases} \tag{23}$$

where $\beta, \mu \succ 0$ with β and μshowcases the scale and shift parameter. A Lévy flight isestimated as per (24).

$$Lé(x) = 0.01 \times \frac{RN_a \times v}{|RN_b|^{\frac{1}{\gamma}}} \tag{24}$$

where RN_a and RN_b denote 2 distributed random values from [0,1]. refers a constant, and v can be measured as per (25) with $(g) = g - 1$ (with 'g' as certain value).

$$v = \left(\frac{(1+) \times \sin\left(\frac{\pi\gamma}{2}\right)}{\Gamma\left(\frac{1+\gamma}{2}\right) \times \times 2^{\left(\frac{\gamma-1}{2}\right)}}\right)^{\frac{1}{\gamma}}$$

3. Results and Discussions

3.1 Dataset Description

Utilizing the process shown in figure 1 the initial experimental setup is created. In this study, CDC's heart disease dataset [23] is used for evaluating the proposed feature selection model. Table 1 describes the total number of instances and attributes in the datasets. The data sample, which includes 319 thousand patients based on 18 criteria, is quite instructive. Figure 2 depicts the distribution of class label of the datasets. It shows the highly imbalance of the dataset. The target variable contains around 250k instances of 'No' variable where the 'yes' variable is less than 50k. The proposed feature selection is evaluated with the following machine learning classifier models: Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF) and Extreme Gradient Boost (XGBoost).

Dataset	No. of Instances	No. of Attributes
Heart Disease	319795	18

Table 1. Dataset Description.

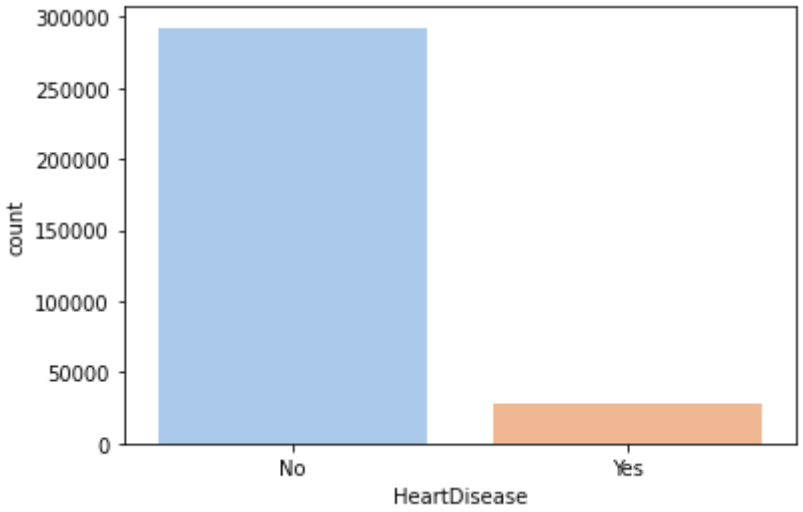


Figure 2. Distribution of Class Label

The feature selection outcome of the MRISSA model is examined in Table 2 with comparative ISSA model in terms of best cost. On the used dataset, the MRISSA model has proven to be successful in selecting the best features at the lowest possible cost. For the taken dataset, the MRISSA and ISSA models have obtained 0.899 and 0.905, respectively, whereas the ISSA model has acquired a lower best cost. The proposed model shown 0.006 increase in the best cost. Since, it is minimal, but it shows an improvement.

Methods	Best Cost	Selected Features
ISSA	0.899	2,3,4,6,7,8,9,13,15
MRISSA	0.905	1,2,3,11,12,14,15,16

Table 2. Selected Features using MRISSA and ISSA algorithm.

The results of classification models for Heart disease dataset are represented in figure 3, where XGBoost had the greatest precision, recall, F1-score and accuracy values, with regard to 0.791, 0.785, 0.788 and 0.784. While SVM's precision, recall, F1-score and accuracy values were the lowest 0.768, 0.763, 0.766 and 0.76101083 respectively. NB attained greater than SVM, 0.774, 0.769, 0.771 and 0.766 with regard to precision, recall, F1-score and accuracy. DT attained 0.780, 0.774, 0.777 and 0.772 with regard to precision, recall, F1-score and accuracy. Although RF's values are greater than DT, it also reported the nearest accuracy to the XGBoost. It demonstrated an average level of performance with values for precision, recall, F1-score and accuracy that were around 0.785, 0.780, 0.783 and 0.778, respectively.

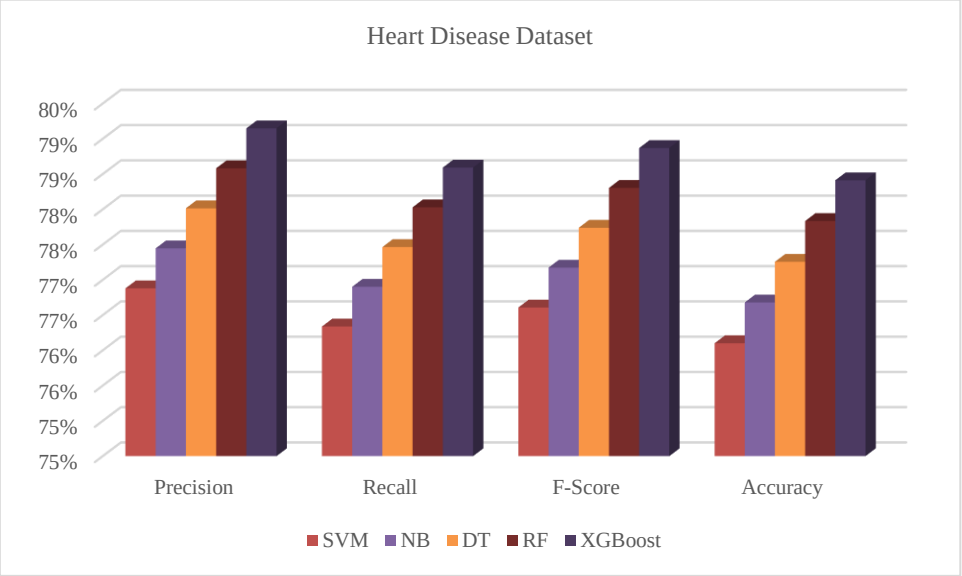


Figure 3 Comparative Results of Heart Disease Dataset

4. Conclusion

A framework for nature inspired optimization algorithm with machine learning classifiers is presented in this study. On the publicly available heart disease datasets, the performance of the suggested algorithm has been examined and assessed. The findings demonstrated that the characteristics chosen using MRISSA significantly improved accuracy in comparison to other techniques. Based on evaluation results, it can be said that the heart-disease dataset showed good performance from both RF and XGBoost models in terms evaluation metrics and MRISSA-XGBoost produced better results. Additionally, it could be inferred that the taken dataset has a problem with class imbalance in addition to the existence of strongly correlated features. Therefore, it is necessary to designed a quick and effective class imbalance technique to improve the accuracy of prediction.

References

[1] Thong NT (2015a) Intuitionistic fuzzy recommender systems: an efective tool for medical diagnosis. Knowl Based Syst 74:133–150
[2] Chen L, Li X, Yang Y, Kurniawati H, Sheng QZ, Hu HY, Huang N (2016) Personal health indexing based on medical examinations: a data mining approach. Decis Support Syst 81:54–65
[3] Ye J (2015) Improved cosine similarity measures of simplifed neutrosophic sets for medical diagnoses. Artif Intelli Med 63(3):171–179
[4] Wójtowicz A, Patryk Ż, Anna S, Krzysztof D (2016) Solving the problem of incomplete data in medical diagnosis via interval modeling. Appl Soft Comput 47:424– 437
[5] Ye J, Jing F (2016) Multi-period medical diagnosis method using a single valued neutrosophic similarity measure based on tangent function. Comput Methods Progr Biomed 123:142–149
[6] AlMuhaideb S, Menai ME (2016) An individualized preprocessing for medical data classifcation. Proced Comput Sci 82:35–42
[7] Leonard M, O'Connell H, Williams O, Awan F, Exton C, O'Connor M, Adamis D, Dunne C, Cullen W, Meagher DJ (2016) Attention, vigilance and visuospatial function in hospitalized elderly medical patients: Relationship to neurocognitive diagnosis. J Psychosom Res 90:84–90
[8] Kurzyński M, Majak M, Żołnierek A (2016) Multiclassifer systems applied to the computer-aided sequential medical diagnosis. Biocybern Biomed Eng 36(4):619–625
[9] Bruijne M (2016) Machine learning approaches in medical image analysis: from detection to diagnosis. Med Image Anal 33:94–97
[10] Thong NT (2015b) HIFCF: an efective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis. Expert Syst Appl 42(7):3682–4370
[11] Meesad, P., and Yen, G. G. (2003). Combined numerical and linguistic knowledge representation and its application to medical diagnosis. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 33, 2, 206–222.
[12] Chabat, F., Hansell, D. M., and Yang, G. (2000). Computerized decision support in medical imaging. IEEE Engineering in Medicine and Biology Magazine, 19, 5, 89–96.
[13] Tsipouras, M. G., Voglis, C., and Fotiadis, D. I. (2007). A Framework for Fuzzy Expert System Creation—Application to Cardiovascular Diseases. IEEE Transactions on Biomedical Engineering, 54, 11, 2089–2105.
[14] Kwiatkowska, M., Atkins, M. S., Ayas, N. T., and Ryan, C. F. (2007). Knowledge- Based Data Analysis: First Step Toward the Creation of Clinical Prediction Rules Using a New Typicality Measure. IEEE Transactions on Information Technology in Biomedicine, 11, 6, 651–660.
[15] Isola, R., Carvalho, R., and Tripathy, A. K. (2012). Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN. IEEE Transactions on Information Technology in Biomedicine, 16, 6, 1287–1295.
[16] Çomak, E., Polat, K., Güneş, S., and Arslan, A. (2007). A new medical decision making system: Least square support vector machine (LSSVM) with Fuzzy Weighting Pre-processing. Expert Systems with Applications, 32, 2, 409–414.
[17] Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications, 38, 4, 4600–4607.
[18] Kahramanli, H. and Allahverdi, N. (2009). Extracting rules for classification problems: AIS based approach. Expert Systems with

Applications, 36, 7, 10494– 10502.
[19] Sekar, B. D., Dong, M.-C., Shi, J. and Hu, X. Y. (2012). Fused Hierarchical Neural Networks for Cardiovascular Disease Diagnosis. IEEE Sensors Journal, 12, 3, 644–650.
[20] Markowska-Kaczmar, U. and Matkowski, R. (2006). Experimental Study of Evolutionary Based Method of Rule Extraction from Neural Networks in Medical Data. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, Lecture Notes in Computer Science, 4065, 76–90.
[21] Mosavi, M.R. and Khishe, M., 2017. Training a feed-forward neural network using particle swarm optimizer with autonomous groups for sonar target classification. Journal of Circuits, Systems and Computers, 26(11), p.1750185.
[22] Jain, M., V. Singh, and A. Rani. 2019. A novel nature-inspired algorithm for optimization: Squirrel search algorithm. Swarm and Evolutionary Computation 44 (2):148–75. doi:10.1016/j.swevo.2018.02.013.
[23] Internet source as on dated 27-Sep-2022 https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease