Tech Science Press

Cuckoo Optimized Convolution Support Vector Machine for Big Health Data Processing

Eatedal Alabdulkreem[1], Jaber S. Alzahrani[2], Majdy M. Eltahir[3], Abdullah Mohamed[4], Manar Ahmed Hamza[5,*], Abdelwahed Motwakel[5], Mohamed I. Eldesouki[6] and Mohammed Rizwanullah[5]

[1]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, 11671, Saudi Arabia
[2]Department of Industrial Engineering, College of Engineering at Alqunfudah, Umm Al-Qura University, Mecca, 24382, Saudi Arabia
[3]Department of Information Systems, College of Science & Art at Mahayil, King Khalid University, Abha 62529, Saudi Arabia
[4]Research Centre, Future University in Egypt, New Cairo, 11845, Egypt
[5]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam Bin Abdulaziz University, AlKharj, 16278, Saudi Arabia
[6]Department of Information System, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, AlKharj, 16278, Saudi Arabia
*Corresponding Author: Manar Ahmed Hamza. Email: ma.hamza@psau.edu.sa
Received: 12 March 2022; Accepted: 22 April 2022

Abstract: Big health data collection and storing for further analysis is a challenging task because this knowledge is big and has many features. Several cloud-based IoT health providers have been described in the literature previously. Furthermore, there are a number of issues related to time consumed and overall network performance when it comes to big data information. In the existing method, less performed optimization algorithms were used for optimizing the data. In the proposed method, the Chaotic Cuckoo Optimization algorithm was used for feature selection, and Convolutional Support Vector Machine (CSVM) was used. The research presents a method for analyzing healthcare information that uses in future prediction. The major goal is to take a variety of data while improving efficiency and minimizing process time. The suggested method employs a hybrid method that is divided into two stages. In the first stage, it reduces the features by using the Chaotic Cuckoo Optimization algorithm with Levy flight, opposition-based learning, and distributor operator. In the second stage, CSVM is used which combines the benefits of convolutional neural network (CNN) and SVM. The CSVM modifies CNN's convolution product to learn hidden deep inside data sources. For improved economic flexibility, greater protection, greater analytics with confidentiality, and lower operating cost, the suggested approach is built on fog computing. Overall results of the experiments show that the suggested method can minimize the number of features in the datasets, enhances the accuracy by 82%, and decrease the time of the process.

1 Introduction

Several medical types of equipment have already been integrated using detectors to gather, exchange, and combine the huge amounts of medical information created. Wireless Sensor Networks (WSN), as well as the Internet of Things (IoT), are examples of developing technologies used in advanced medical systems (IoT). Furthermore, new mobility projects are being widely implemented, which is accelerating the evolution of smart medical systems. The goal is to make the most of actual data coming from a variety of health and sensory offerings. The Internet of Things creates a wide range of sophisticated large health records. Both storing and processing systems face numerous issues as a result of this information. To overcome difficulties, the integration of IoT with several key technologies, such as cloud technology, is becoming crucial.

Organizations can collect, collect, organize, and manipulate enormous amounts of information from a combination of disciplines. At the optimum speed and also at the correct time to obtain new perspectives [1]. Different aspects of the solution process can be ineffective or lacking in knowledge. Due to the existence of noise, a few others may confuse the classifier's cost and effectiveness in the prediction performance. While leaving such measures put in place causes no data concerns, it enhances the computational time and slows down the system's reaction times. Because of the limited storage capacity, this also leads to the collection of a quantity of non-useful information alongside beneficial information. Furthermore, the individual feature selection approach and the particular issue may be linked, making the method inapplicable to certain other situations [2].

Artificial neural networks (ANNs) are used in machine learning to create neurons with trainable weights so that the neuron, has a relevant mathematical functionality. It is related to the information in the database as feasible [3]. Convolution neural networks (CNNs) for prolonged spatial types of data (e.g., voice and image validation) [4,5], recurrent neural networks (RNNs) for statistical kinds [6], and generative adversarial networks (GANs) for producing new instances and categorizing instances [7] are all instances of deep learning methods. For massive amounts of data, deep learning is just an effective and simple data-mining technique. Moreover, for enhanced quality, deep learning depends on effective equipment, particularly the graphics processing unit (GPU), and such equipment has been costly [8,9].

By picking characteristics and categorizing the data, the SVM seems to be another well-known and successful supervised learning algorithm. Leading to the formation of deep learning, SVMs significantly outperforms ANNs in a variety of real-world uses. That includes healthcare, the technology sector, online analysis, spectroscopic unsupervised feature pixel density, extremely unbalanced sets of data, financially distressed extraction, clustering techniques, etc. In contrast to deep learning methods that attempt to be based on observations using ANNs. Then the SVM uses mathematical optimization to divide (rather than combine) various classes of information using kernels. Furthermore, an SVM offers high precision with low computational resources and limited information, which are 2 of deep learning's drawbacks [10–12].

The main contribution of the proposed method is given below:

- By quickly analyzing the medical information in real-time by boosting correctness while reducing computing expenses.

- It can be accomplished by using the Hybrid Chaotic Cuckoo Optimization algorithm used for feature selection and Convolutional Support Vector Machine (CSVM)
- In an Ambient Healthcare method to conduct feature selection on information and then categorize it to decrease computation time and enhance the performance.
- To minimize the randomized feature selection method and reduce becoming caught in locally optimal, use chaotic theories, levy flight, and the disruption operator.
- Opposition-based training stops ideas from clumping together in one location, allowing for a more thorough exploration of the issue space.
- Because the disruption operators examine the opposing points of answers, the responses span more of the issue surface and produce better outcomes.

The rest of our research article is written as follows: Section 2 discusses the related work on Healthcare systems, feature selection methods, and several classifiers. Section 3 shows the algorithm process and general working methodology of the proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2 Literature Review

Health services should expect considerable breakthroughs as a result of this approach of IoT and cloud technology. This connection allows for the development of new types of smart medical devices. Medical systems that have lately been created and launched are aimed at practitioners and academicians in order to adopt contemporary health services. Healthcare information is recorded, processed, and analyzed using IoT-based healthcare infrastructure. In this spirit, over the last few times, creating medical systems, extraction of features, and clustering techniques has gotten a lot of interest in industry and academia [13]. A current healthcare design will be described in detail in the following.

Feature extraction is a critical pretreatment and information retrieval tool for information analytics and it is used in a wide range of fields. It is often used as a key method for high-dimensional research methodology. Since it has demonstrated its potential to improve a wide range of many other academic tasks, including clustering techniques, content analysis, image recognition, and video data, to mention a few [14,15]. Features extraction troubles had a good compromise between both the volume of extracted features and the reliability of the categorization. In reality, the overall length of the features subdomain was frequently predetermined in prior investigations.

Classification approaches can be categorized as follows based on how to describe the characteristics is used in multiple databases trained, semi-supervised, and unstructured [16]. The information that describes the characteristics is used to assist the features extraction procedure in supervised instructional methods. Such techniques use labeled data for training feature selection algorithms, which estimate the importance and relevance of characteristics according to specified criteria [17,18]. Unsupervised techniques are effective for picking characteristics with regard to the labeling in databases. Since unlabeled training comprises examples and characteristics with no knowledge about the normal categorization of observations [19]. Mixed methods merge multiple different algorithms, namely filtering and wrappers methodologies, to produce much more satisfying results for finding solutions [20].

For conserving effort and lowering processor expenses, issues of velocity, economy, and computation complexity can be tackled. Designers have to lower the volume of information being analyzed, which we may do by minimizing the features of big data processing. A Feature Selection (FS) strategy

can be used to reduce the amount of data that is stored. FS has an impact on the results and allows for limited circumstances. The characteristic that is being used to enhance efficiency is determined by FS [21].

The program characteristics and hyperparameters are optimized using metaheuristic optimization techniques [22]. Chaos is just a predictable part of an emerging whose input variables and settings affect its behavior. The essence of chaos appears to be random, but it also has a consistency to it. Chaos is a non-linear function that can alter the behavior of a program. Chaotic systems describe a non-converging limited pseudo-random controlled system. Minimizing local minimum standards can address standards developed in optimization computation. There seem to be a lot of maps that are chaotic [23–25].

Selecting an optimum information domain is crucial for speeding up and grading the categorization, particularly when dealing with huge amounts of information. The cuckoo's method, which also was developed in this work, was utilized to pick characteristics. Cuckoos are fascinating species that use a spatial approach to reproduce aggressively. They really lay hundreds of eggs of many other birds known as hosts. If the hosting birds locate the cuckoo eggs in the nests and push those away, they would not have an opportunity to make a difference. If indeed the hosting birds are unable to locate the cuckoo's egg, the cuckoo's chicken, who develops quicker than the hosting chicken, would evict them from the nests [26].

3 Proposed Health Data Methodology

The stages of the suggested strategy are organized in a step-by-step manner. The first step begins with data collected from different sources. At this point, data variability is a problem. Information is transported to the second step in order to complete the data management process. The information is stored in the second stage, then optimized and categorized in a manner that allows the third stage to perform well and offer flawless solutions. The steps of the CCO-LFD-CSVM approach will be described in depth in the following sections.

Fig. 1 shows the proposed workflow. It consists of three phases:

- Data Collection
- Data management
- Feature Selection and Classification

The detailed descriptions were explained below.

3.1 Ambient Intelligent Healthcare System Based Big Data

Throughout many powerful data analysis scenarios, information changes throughout the period and should be examined in near real-time. Models developed to interpret such data soon grow outdated as structures and correlations develop and mature. In statistics and machine learning, this is referred to as notion drifting. Concept drift is a term used in machine learning and data mining to explain variations in the connections among a set of data in an issue across periods. When it comes to interacting with idea drift, the most typical strategy is to ignore it and assume that information somehow doesn't alter.
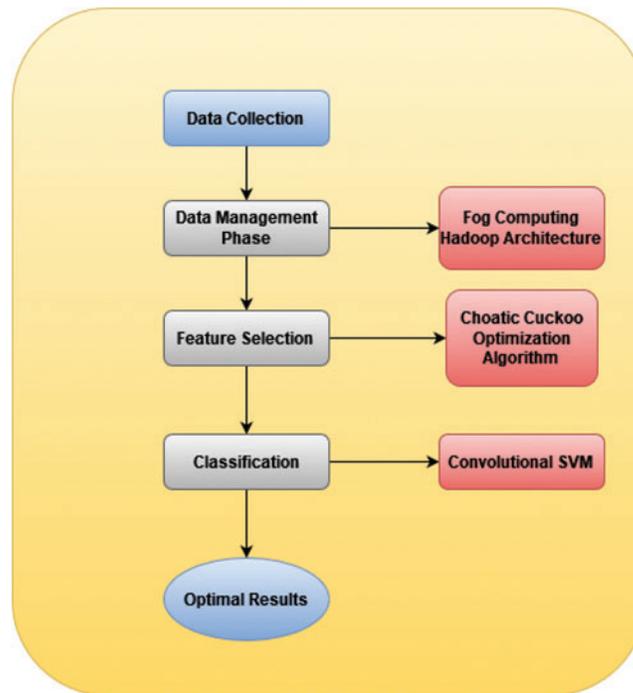
Figure 1: Proposed workflow of CCO-LFD-CSVM

It's critical to solve the issue of long production times and high computing costs in healthcare big data systems. This can be accomplished by (i) presenting a method for analyzing different forms of medical data, (ii) forecasting meaningful information with minimal computational expense, and (iii) analyzing statistical information. As a result, a hybrid method with 2 stages is developed. To begin, the set of attributes is reduced using a feature selection method. After then, the suggested hybridization algorithm's second step is data categorization. Fig. 2 shows the Architecture of Ambient Healthcare System based Big Data.

3.1.1 Data Collection

Each stage is divided into two parts: one is for interpretation of data and the other for transmitting gathered information to another stage. The information originates from a variety of places, including clinics, research facilities, wearable technology, and government agencies. The acquired information is then transmitted to the next step through a communication channel.

3.1.2 Data Management Phase

Reduced latencies and real-time communications during the data analysis stage and other phases are provided via fog technologies. This stage is carried out with the help of Hadoop [1], an open-source Java-based software platform. Hadoop's fundamental goal is to spread data warehouses and software execution across huge groups. Hadoop enables vast capacity for just about any type of information via the Hadoop Distributed File System (HDFS). The information management stage is divided into two parts, each of which is in charge of storage and analysis. The first component is information processing which uses the HDFS [1] to store information. Its developed system is information processing and

categorization, which employs Hadoop MapReduce [1] computing and simultaneous calculating to analyze a variety of data using the suggested hybrid method.
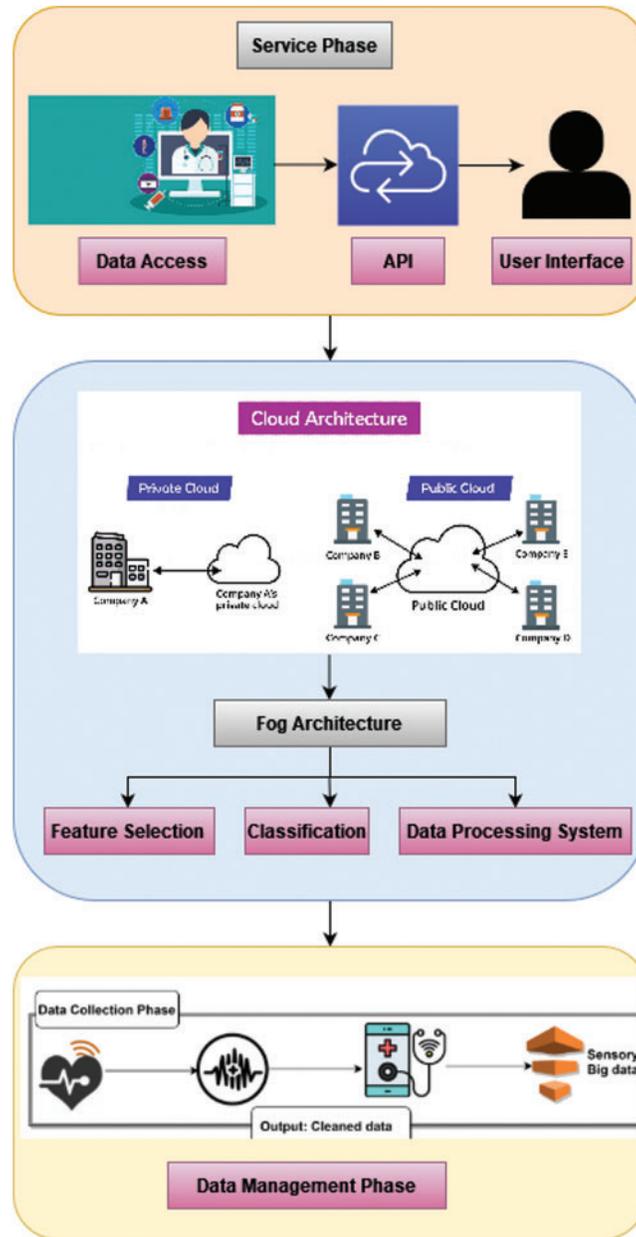
Figure 2: Architecture of ambient healthcare systems based on big data

This phase's execution entails employing a mixture model to optimize data. This approach uses Chaotic Cuckoo Optimization to choose features from big data stored in HDFS. Then uses Convolutional SVM to classify the optimized data, and all of this is done using MapReduce computing and parallel computation, as illustrated in Fig. 3.
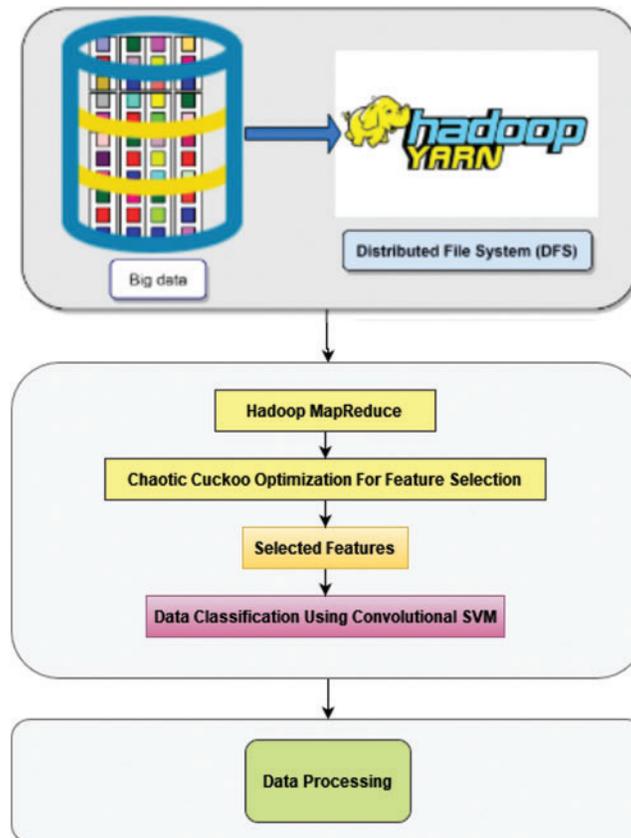
Figure 3: Data processing using the proposed method

3.1.3 Service Phase

The data processing, Application Programming Interface (API), and User Interface (UI) components make up the overall stage. Those modules communicate with one another in order to make the best decisions possible. The shared data component takes information and facts from the detection and analysis module and prepares them for usage by the API and UI components.

3.2 Chaotic Cuckoo Optimization Based on Levy Flight and Disruption Operator

As a feature selection approach, the suggested wrapper-based chaotic cuckoo optimization algorithm with levy flight and disruption operator (CCO-LFD) is discussed. The cuckoos wander around in the solution space to change their views to any location in the continuous domain within the basic cuckoo optimization procedure. Its answers are restricted to the binaries field [0,1] numbers given the nature of both the feature selection challenges.

Matrices can be created with NM_{pop} cuckoos $NM_{pop} \times NM_{var}$. In the traditional technique, cuckoos randomly chose certain shelters which are established in the process at first sight of laying eggs. The chaotic functional accelerates the speed of convergence. Utilizing ergodicity, smoothness, and semi-stochastic qualities to seek the best solution in a chaos method amongst local optimal solutions led to the high probability of optimal solutions. Chaos factors, on the other hand, now are supposed to be

produced by an unstructured one-dimensional map that allows them to change in the region of (0.1) or be scaled in this region.

A random number is used as the beginning value for the Gauss/mouse chaotic functions, and the serial numbers are generated using Eq. (1)

$X_o = Rand$

$$X_{i+1} = \begin{cases} 1 & X_i = 0 \\ \dfrac{1}{mod(X_i, 1)} & Otherwise \end{cases} \tag{1}$$

Cuckoos would look about for the ideal habitat in an attempt to discover it for the next generations. Despite the fact that there are numerous evolving variations in the research, cuckoo optimization method difficulties such as convergence rate and ineffective outputs remain. To overcome these challenges, the Levy flights approach is utilized, which allows the CCO-LFD method to produce more accurate alternatives. In addition, the disruption operators detected habitat feature subset. Cuckoo optimization algorithm (COA), it is unable to execute full search properly, will be able to do it more successfully and will not be locked in local minima to use this strategy.

The step size is then computed using Eq. (2). The duration scaling of 0.01 is used to reduce inflating the amount of the steps and causing the solutions to bounce out of the modeling environment.

$Step\ size = 0.01 \times Ss \tag{2}$

The percentage the number of eggs lay through this cuckoo and the overall amount of eggs in limit determines the egg-laying radius (ELR) of each cuckoo. an is adjusting ELR's highest benefit. ELR is written as Eq. (3).

$$ELR = \alpha \times \frac{Number\ of\ Current\ Cuckoos\ eggs}{Total\ amount\ of\ eggs} \times (Var_{High} - Var_{Low}) \tag{3}$$

In Eq. (4), the universal randomized searching to use the idea of levy flights is provided; where $X_i^{(t)}$ represent the current situation of the cuckoo I after t repetitions during its last s a home. To construct a new phase $X_i^{(t+1)}$ of $X_i^{(t)}$ Levy flight is used. Eq. (4) calculates the curr of the habitats $X_i^{(t)}$ via levy flight, and ELR is specified in Eq. (3).

$X_i^{(t+1)} = X_i^{(t)} + ELR \times Levy(Ss) \tag{4}$

Because the continuously collection of characteristics must be translated to the matching binaries field, the very next equation uses a new location of a habitats, where $Y^{j,t+1}$ makes reference to the cuckoo with the j index upon t iterative process:

$$Y^{j,t+1} = f(X) = \begin{cases} 1 & if\ (W(X^{j,t+1})) \geq Rand \\ 0 & Otherwise \end{cases} \tag{5}$$

$W(X^{j,t+1})$ Is measured by using Eq. (6)

$$W(a) = \frac{1}{1 + e^{10(a-0.5)}} \tag{6}$$

The constant field (free location) should be translated into binary methods to fix the feature selection challenges.

When the result of a bit exceeds 1, the relevant feature inside the extracted features is picked, whereas 0 signifies zero. It signifies that the feature subset's 2nd and 4th features have been chosen. Eq. (7) describes the fitness value:

$$fitness_t = \max\left((1 - b) \times ACC + b \times \left(1 - \frac{M}{N}\right)\right) \tag{7}$$

Fitness values are set to the cuckoo's habitats in the search process throughout each repetition. Every repetition evaluates these locations, with the ideal position being chosen as the optimal answer. ACC is defined as accuracy.

In principle, the suggested approach relies on the chaotic method, levy flight, opposition-based training, and disruption operators to improve the behaviors of the cuckoo optimization technique. Obviously, both of these operatives have their own set of responsibilities. The goal of employing the chaotic algorithms in the first phase is to enhance the starting spot. The cuckoos get the greatest impact on the converging of answers resulting to the comprehensive solution; hence the updated method was applied in the levy fly operation. In addition, opposition-based training is a clever way to explore the other side the space and improve exploring strength.

Two factors are chosen to end the process in the CCO-LFD method: maximal repetition (or creation) in the last 20 iterations, there has been no substantial improvement in average accuracy.

Algorithm 1: CCO-LFD Algorithm

Input: NM$_{pop}$, MaxGeneration=100, NM$_{var}$, Cuckoos Num=50, t=1
Output: Ss$_{Best}$
Step 1: Begin population
Step 2: For initial habitats select the chaotic map
Step 3: X$_0$=rand
Step 4: for i=0: NMpop
Step 5: Generation of NM$_{pop}$ outcomes and the solution is given into the populations.
Step 6: Levy Flights is called by using $Ss = \dfrac{u}{|v|^{\frac{1}{\beta}}}$

Step 7: Based on the radius of laying eggs, every cuckoo choose the habitats for egg laying
$$habbitat = parentposition + ELR * levy\,(Ss)$$
Step 8: Every position habitat is measured by $habitat = u + l - habitat$
Step 9: Habitats are measured.
Step 10: Compare and choose the best habitat
Step 11: Eliminate the worst quality parents and construct the new ones.
Step 12: Estimate the fitness function for every cuckoo
Step 13: Assigning the rank based on evaluating sort of current population
Step 14: End
Step 15: Keep the optimal habitats
Step 16: Save the current best solution SS$_{best}$
Step 17: Return SS$_{best}$

Algorithm 1 displays the CCO-LFD pseudo-code. The method begins by choosing randomised populations of chaotic equation answers, after which repetition begins.

The primary goal of each iteration is to find the optimal solution, followed by the factors. During the flocking phase, levy flight is used to modify the population's answers, and the reverse location of every nesting is estimated as fresh nests. It then uses the dop algorithm to create better appropriate placements for the entire habitat and their opponents. This approach is continued until a satisfactory ending restriction is found. The method then returned to the better outcome. Fig. 4 shows the Framework of Proposed Method.
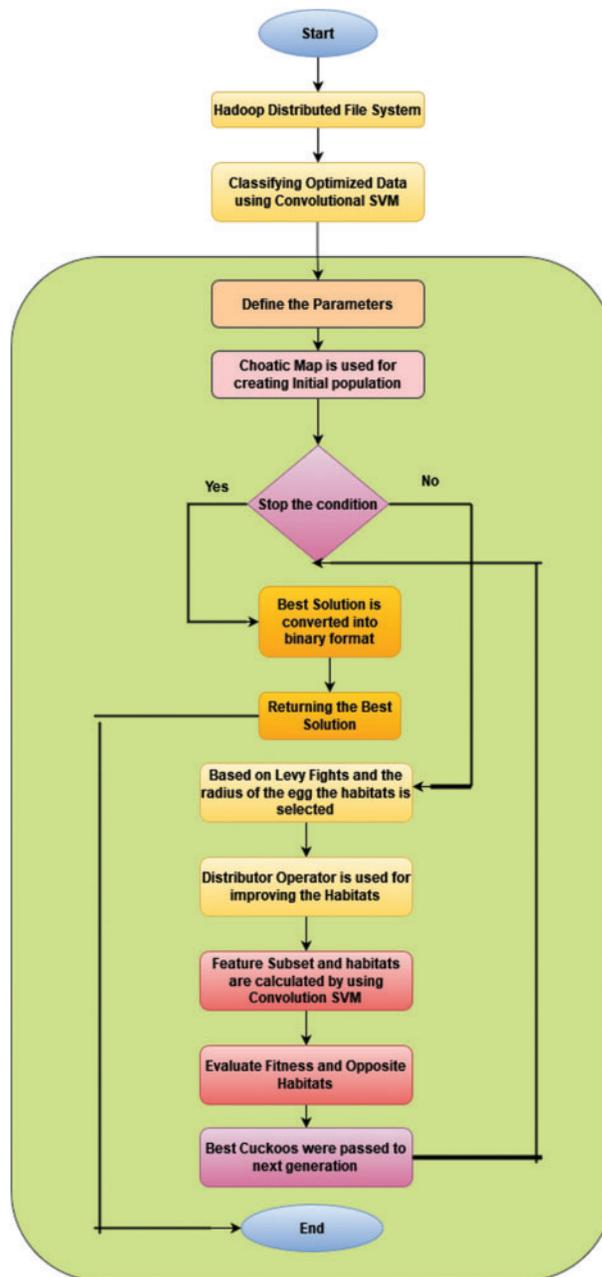
Figure 4: Proposed classification model using convolutional SVM (CSVM)

The main distinction among the suggested CSVM and the existing SVM is the convolution products. Consequently, the Convolution products, including classic and proposed new convolution products.

3.2.1 Convolution Process

CNN is one of the most important deep learning methods, and its effectiveness has indeed been validated in a variety of recognized areas of study. We present a few of the critical CNN operating methods which are used in research work.

1. Padding: To avoid the information reduced size caused by convolution process in next level, we append 0 around in the source images, a procedure known as padding.
2. Stride: A stride is a kernel that moves an up or down length each cycle. The larger the stride, more the autonomous the convolution process' surrounding results are.
3. Convolution: Multiply of said data among the inputs as well as the kernel (filter) goes through each action of convolution depending on the provided duration after padded. These goods are added together and placed in the appropriate spots on next level.

3.2.2 Repeated Attributes Based Convolution Product

Assume that amounts of inferences, answers, and filters produced in each answer, and factors included within every filtering are N_{att}, N_{sol}, N_{filter}, and N_{var}, correspondingly. Let $v_{r,a} = v_{r,a,0}$ be the value of the ath attribute in the rth record, and $v_{r,a,f}$ be the value of the ath attribute in the rth record after applying the fth filter, where a $1, 2, \ldots, N_{att}$, s $1, 2, \ldots, N_{sol}$, and f $1, 2, \ldots, N_{filter.}$

Algorithm 2: CSVM Algorithm

Input: Dataset used

Output: The classifier CSVM producing a optimum accuracy

Step 0: Randomly divide the datasets into k folds and choose the one-fold, it has N tires

Step 1: Implementation of CSVM (i,k∗) using ith parameter and k∗th fold for dataset i=1,2,...,N and it finds the maximum accuracy between all N tries.

Step 2: Implementation of CSVM (i∗,j) using jth fold dataset by i∗th try j=1,2,...,k

Step 3: Procedure CSVM (α, β)

Step W0: Randomly select the solution X and estimate $F(X_{s,f})$ using convolution product and SVM.

Step W1: Assign t=1

Step W2: Randomly selected solutions were updated.

Step W3: t is maximized by 1 therefore t=t+1 and then it follows Step W2 if t<N_{gen}

Step W4: End, $F(X_{gBest}$, pFilter[$_{gBest}$]) is the accuracy.

The suggested CSVM's pseudo - code, as well as the integrating of the suggested convolution product d, are explained in Algorithm 2.

4 Result Analysis

The proposed ambient health care architecture processing using CCOLFDO-CSVM method uses evaluation metrics such as accuracy, computation time and varieties of data.

4.1 Datasets

The samples that have been gathered from Kaggle [1] are examined in this section. Those databases would be used to validate the method and predict the outcome. As noted in the first phase's descriptions, they come in a variety of sorts, and the suggested technique would easily accommodate them. Tab. 1 the properties of the datasets used are summarized.

Table 1: Datasets used

Dataset used	No of instances	Features used	Type of classification	Datasets availability
Heart attack prediction	305	82	Multi-class	https://www.kaggle.com/imnikhilanand/heart-attack-prediction
Pima Indians diabetes	774	11	Binary class	https://www.kaggle.com/uciml/%20pima-indians-diabetes-database
Heart disease UCI	315	17	Multi-class	https://www.kaggle.com/ronitf/heart-disease-uci
Sonar	1340	63	Binary class	https://www.kaggle.com/ypzhangsam/sonaralldata

4.2 Comparison of Algorithms

The findings of evaluating the hybrid CCO-LFD-CSVM method are presented in this Ambient health care approach. Firstly, we'll use two methods to compare the speed and accuracy of execution for each dataset we've examined. First is to use simply CSVM to perform categorization. The second approach is to use CCO-LFD and CSVM to do feature selection and classifications, which is the combination method discussed before. It is introduced first-way findings that provide CSVM findings on numerous datasets without using any additional methods. After running the suggested hybrid method, the second-way outcomes would be calculated. Tab. 2 will display the comparative outcomes. The actual and anticipated information forms for several samples are depicted in Fig. 5.

Table 2: Comparison of accuracy and speed

Classifiers Used		Datasets Used			
Algorithms	Parameters	Diabetes	Sonar	Heart-C	Heart-H
CSVM	Total Features	8 of 8	60 of 60	13 of 13	13 of 13
	Accuracy (%)	78.32	86	84.01	84.45
	Time (s)	1.3042	0.86015	0.80142	0.81422
CCOLFDO-CSVM	Total Features	4 of 8	52 of 60	12 of 13	12 of 13
	Accuracy (%)	82.35	89.21	86.24	88
	Time (s)	0.9231	0.8274	0.7952	0.7745
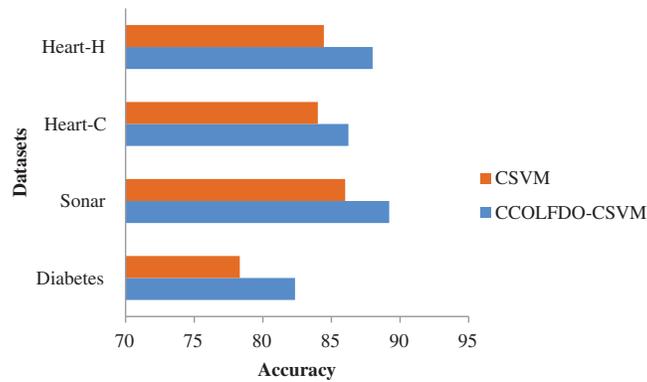
Figure 5: Comparison of accuracy

The proposed CCO-LFD-CSVM method achieves better accuracy and time in all the datasets. In diabetes it achieves 82.35%, Sonar it achieves 89.21%, Heart-C it achieves 86.24% and Heart-H it achieves 88% of accuracy. In diabetes it takes 0.9231 s, Sonar it takes 0.2874 s, Heart-C it takes 0.7952 s and Heart-H it takes 0.7745 s of times.

Equations can be used to compute the Classification Performance or Accuracy in Eq. (8). Tab. 3 presents the results of the suggested CCO-LFD-CSVM method's confusion matrix.

Precision, as shown in Eq. (9), indicates how often real positive samples were found among all positive projected samples. Eq. (10) shows us how many test results are identified out of the total number of positive samples.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8}$$

$$Precision = \frac{(TP)}{(TP + FP)} \tag{9}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{10}$$

Sensitivity is a ratio of positive data points that is accurately regarded positive to all positive datasets and is determined with Eq. (11).

$$Sensitivity = \frac{TP}{TP + TN} \quad (11)$$

Specificity is a proportion of negative data points that is wrongly regarded as positive when compared to all negative datasets. Equation can be used to compute it (12).

$$Specificity = \frac{TN}{FP + TN} \quad (12)$$

Table 3: Results of confusion matrix

Metrics	TP	FP	FN	TN	Precision	Recall	Specificity	Sensitivity
Diabetes	4860	430	1150	1450	93%	82.66%	79%	83.41%
Sonar	995	135	113	890	89%	92.13%	89.3%	93.74%
Heart Attack-C	1120	270	188	1480	82.5%	87.96%	87%	87.45%
Heart Attack-H	1680	260	135	1480	84%	90.98%	79.65%	94.10%

The receiver operating curve (ROC) diagram is a representation of the differential among the true-positive and false-positive rates of classifiers. This enables a method of determining which classifiers are on average better under the ROC curve (AUC). Figs. 6–9 shows the ROC Curve increases the accuracy of the dataset using CSVM classifier.
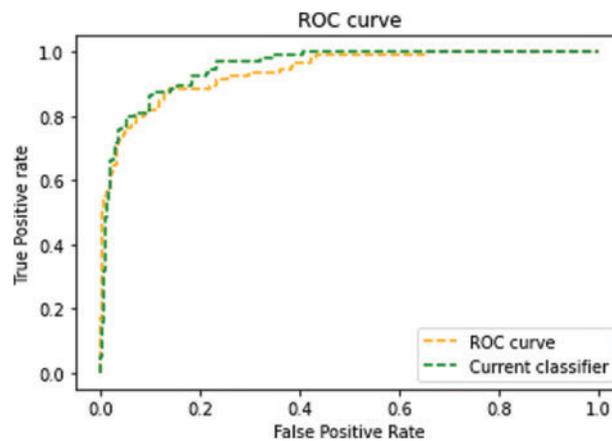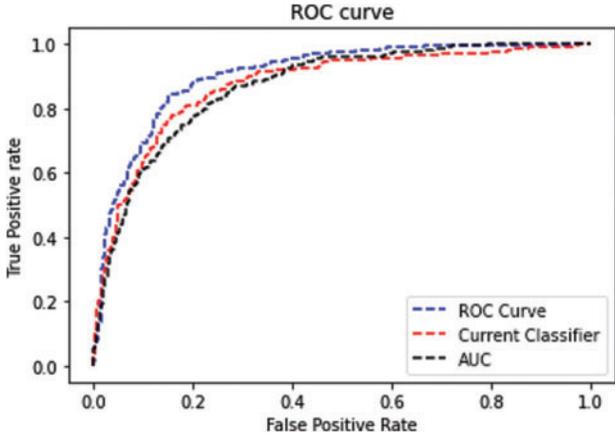

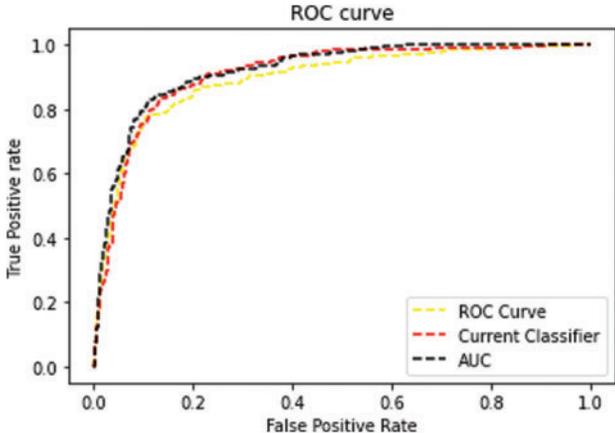**Figure 6:** The ROC curve: diabetes

Figure 7: ROC Curve: Sonar
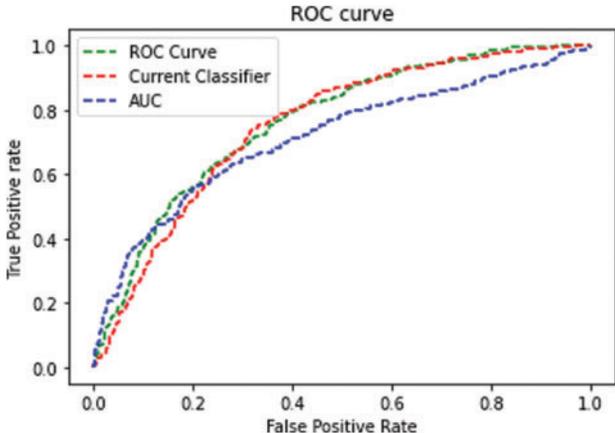

Figure 8: ROC Curve: Heart-C

Figure 9: ROC Curve: Heart-H

5 Conclusion

Several medical big data projects require just too much energy to provide mankind with relevant knowledge that can assist improve and grow this profession at an affordable price. As a result, a hybrid algorithm which is based on diverse forms of health data. It will be simple for everyone to forecast information and offer important data and details in shall submit it to a number of groups who are interested. There are two processing steps in the CCO-LFD-CSVM method. This is to optimize data in order for the second one to be more efficient, with the latter one being in charge of categorizing the optimized data. The suggested technique improves and boosts efficiency by around 6%. This can improve processing capability by decreasing the time required to process data by about 7%. The CCO-LFD method employs feature selection optimization to transform random variables into chaotic behavior. Furthermore, in high-dimensional space, the levy flight is utilized to determining factors and modifies cuckoo's movements. Furthermore particular, the disruption operator and the opposition-based learning method. That employed to boost detection capability and guarantee feature subset in order to expedite the computational efficiency. In future convolutional neural network can be used. The machine learning classifier perform with good accuracy and future ensemble model can be used for testing.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Alwateer, A. M. Almars, K. N. Areed, M. A. Elhosseini, A. Y. Haikal *et al.,* "Ambient healthcare approach with hybrid whale optimization algorithm and naïve Bayes classifier," *Sensors*, vol. 21, pp. 13, 2021.

[2] J. Hamidzadeh, "Feature selection by using chaotic cuckoo optimization algorithm with levy flight, opposition-based learning and disruption operator," *Soft Computing*, vol. 25, no. 4, pp. 2911–2933, 2021.

[3] W. C. Yeh, Y. Jiang, S. Y. Tan and C. Y. Yeh, "A new support vector machine based on convolution product," *Complexity*, vol. 2021, no. 4, pp. 1–19, 2021.

[4] L. Liu, W. Huang and C. Wang, "Hyperspectral image classification with kernel based least squares support vector machines in sum space," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1144–1157, 2018.

[5] B. Liu, Q. Hu, X. Chen and J. You, "Stroke sequence dependent deep convolutional neural network for online handwritten Chinese character recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4637–4648, 2020.

[6] Z. Cui, F. Xue, X. Cai, Y. Cao, G. G. Wang *et al.,* "Detection of malicious code variants based on deep learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3187–3196, 2018.

[7] D. Li, L. Deng, B. Bhooshan Gupta, H. Wang and C. Choi, "A novel CNN based security guaranteed image watermarking generation scenario for smart city applications," *Information Sciences*, vol. 479, no. 2, pp. 432–447, 2019.

[8] S. Mittal and S. Vaishay, "A survey of techniques for optimizing deep learning on GPUs," *Journal of Systems Architecture*, vol. 99, no. 4, p. 101635, 2019.

[9] Y. Sun, B. Xue, M. Zhang and G. Yen, "Completely automated CNN architecture design based on blocks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 4, pp. 1242–1254, 2019.

[10] L. Tang, Y. Tian and P. M. Pardalos, "A novel perspective on multiclass classification: Regular simplex support vector machine," *Information Sciences*, vol. 480, no. 3, pp. 324–338, 2019.

[11] V. Blanco, A. Japón and J. Puerto, "Optimal arrangements of hyperplanes for SVM-based multiclass classification," *Advances in Data Analysis and Classification*, vol. 14, no. 1, pp. 175–199, 2020.

[12] S. C. Tsai, C. H. Chen, Y. T. Shiao, J. S. Ciou and T. N. Wu, "Precision education with statistical learning and deep learning: A case study in Taiwan," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–13, 2020.

[13] S. Yadav, A. Ekbal and S. Saha, "Feature selection for entity extraction from multiple biomedical corpora: A PSO-based approach," *Soft Computing*, vol. 22, no. 20, pp. 6881–6904, 2018.

[14] M. Zhang, F. Chen, B. Tian and D. Liang, "Multi-temporal SAR image classification of coastal plain wetlands using a new feature selection method and random forests," *Remote Sensing Letters*, vol. 10, no. 3, pp. 312–321, 2019.

[15] T. Luo, C. Hou, F. Nie, H. Tao and D. Yi, "Semi supervised feature selection via insensitive sparse regression with application to video semantic recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1943–1956, 2018.

[16] P. Bannigidad and C. Gudada, "Age type identification and recognition of historical Kannada handwritten document images using HOG feature descriptors," in *Proc. Computing, Communication and Signal Processing*, Chennai, India, pp. 1001–1010, 2019.

[17] N. Nayar, S. Ahuja and S. Jain, "Swarm intelligence for feature selection: A review of literature and reflection on future challenges," in *Proc. Advances in Data and Information Sciences*, Japan, pp. 211–221, 2019.

[18] M. Mafarja, I. Aljarah, H. Faris, A. L. Hammouri, M. Ala *et al.,* "Binary grasshopper optimisation algorithm approaches for feature selection problems," *Expert Systems with Applications*, vol. 117, no. 3, pp. 267–286, 2019.

[19] T. Thaher, A. A. Heidari, M. Mafarja, J. S. Dong and S. Mirjalili, "Binary Harris Hawks optimizer for high-dimensional, low sample size feature selection," *Evolutionary Machine Learning Techniques*, vol. 8, pp. 251–272, 2020.

[20] L. Kumar and K. K. Bharti, "An improved BPSO algorithm for feature selection," *Recent Trends in Communication, Computing, and Electronics*, pp. 505–513, 2021.

[21] N. Neggaz, E. H. Houssein and K. Hussain, "An efficient henry gas solubility optimization for feature selection," *Expert System Application*, vol. 152, no. 10, p. 113364, 2020.

[22] G. I. Sayed, A. E. Hassanien and A. T. Azar, "Feature selection via a novel chaotic crow search algorithm," *Neural Computing Application*, vol. 31, no. 1, pp. 1–18, 2019.

[23] P. Shunmugapriya and S. Kanmani, "A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid)," *Swarm Evolution Computing*, vol. 36, no. 1, pp. 27–36, 2017.

[24] J. Hamidzadeh and N. Namaei, "Belief-based chaotic algorithm for support vector data description," *Soft Computing*, vol. 23, pp. 1–26, 2018.

[25] L. Kumar and K. K. Bharti, "An improved BPSO algorithm for feature selection," in *Proc. Recent Trends in Communication, Computing, and Electronics*, India, pp. 505–513, 2019.

[26] X. K. Yang, L. He, D. Qu and W. Q. Zhang, "Semi-supervised minimum redundancy maximum relevance feature selection for audio classification," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 713–739, 2018.