

How to Cite:

Sundaravadivel, T., & Mahalakshmi, V. (2022). Dynamic scale factor based elephant herding optimization with weighted support vector machine for COVID-19 prediction. *International Journal of Health Sciences*, 6(S3), 5932–5943.
<https://doi.org/10.53730/ijhs.v6nS3.7311>

Dynamic scale factor based elephant herding optimization with weighted support vector machine for COVID-19 prediction

T. Sundaravadivel

Research Scholar, Dept. of Computer Science and Engineering, FEAT, Annamalai University, Tamilnadu, India

Dr. V. Mahalakshmi

Assistant Professor, Dept. of Computer Science and Engineering, FEAT, Annamalai University, Tamilnadu, India

Abstract---The rapid spread of Corona virus disease 2019 (COVID-19) makes its detection among the greatest challenges worldwide. A recent report indicates that COVID-19 has been detected in over 1.6 million individuals across several countries. To predict the risk of COVID-19, predictive diagnostic models have been developed to assist clinicians with RT-PCR results. However, the models do not provide the level of accuracy that is possible. To solve this problem, the proposed system designed a Dynamic Scale factor based Elephant Herding Optimization (DSEHO) with Weighted Support Vector Machine (WSVM) approach for covid-19 prediction. The initial analysis includes a dataset of COVID-19 patients. The data should be normalized using a min-max method. Independent Component Analysis (ICA) is introduced to decrease the dimensions. To select the optimal features, a Dynamic Scale factor based Elephant Herding Optimization (DSEHO) is implemented. For the prediction of COVID-19, Weighed Support Vector Machines (WSVM) is used. The proposed system provides improved accuracy, precision, recall, and F-measure over the previous system.

Keywords---weighted support vector machine (WSVM), elephant herding optimization (EHO), COVID-19, accuracy rate.

Introduction

A newly emerging global pandemic, Corona virus disease 2019 (COVID-19), is being sourced as a result of Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2). Corona virus disease in 2019 has severe health consequences. A

total of 102 million infections have been registered worldwide, and hospitalization rates can reach 18% depending on the age group [1]. This issue has led to the deaths of more than 2 million [2] and significant costs to the healthcare system. The public health burden associated with COVID-19 is also increased by the collateral damage caused by an increase in non-communicable diseases [3]. Therefore, the World Health Organization has declared this an emergency of public health. Hospital resources have been severely strained as a result. In addition to those who require intensive medical treatment for COVID-19, hospitals are flooded with suspected COVID-19 cases that do not necessarily need hospitalization. To maximize triage and reduce infection risks, it is imperative to determine whether a patient is likely to be positive for COVID-19 immediately upon admission to a healthcare facility.

A vaccine against it has not yet been approved for human use. The disease is spread more rapidly among people who live close together. Additionally, frequent hand washing prevents the spread of viral infections as well as the spread of diseases. The symptoms of most infections include fever and cough. It is possible to experience sore throats in addition to chest discomfort and the production of sputum [4-5]. It is evident from the name that this condition is primarily respiratory in nature. Approximately two to fourteen days are required for both SARS and MERS (Middle East respiratory syndrome) to develop. Therefore, patients are at a high risk for death.

The scientific literature contains a number of diagnostic prediction models. A dynamic causal model framework for COVID-19 has been presented [6]. In accordance with the model, individuals were divided into four categories: location, infection status, test results, and clinical status. A number of parameters for an infectious disease cannot be accurately estimated due to non-specific symptoms. Recent work discusses two concepts. In the first concept, COVID-19 is diagnosed, while in the second concept, people are predicted to become infected soon. A review of the study indicates that most of the commonly used models are inaccurate and biased. Two statistical methods were utilized to approximate and forecast the division of COVID-19 in India a few years ago. These algorithms include Susceptible-Exposed-Infectious-Recovered (SEIR) and regression models. Using SEIR and regression models at the time, it was estimated that there were 10,453 and 6153 confirmed cases in India, respectively. Machine learning is critical in order to effectively diagnose diseases. The following is a summary of the remainder of this document. In the second section, previously discussed methods for predicting COVID-19 are presented. Section 3 describes Dynamic Scale Factor-based Elephant Herding Optimization using Weighted Support Vector Machines (WSVM). Results are described in section 4. This is followed by a conclusion.

Literature review

Guhathakurata et al (2021) presented a novel way to predict COVID-19 by using Support Vector Machines (SVMs). The inputs are COVID-19 symptoms. Based on a patient's symptoms, it's possible to determine whether they're infected. SVM classifier gets applied to the dataset. Based on the results, there are three levels of infection: not infected, mildly infected, and severely infected. The results of the

experiments suggest that the designed system has a rare chance of predicting cases 87% of the time [9]. Khandy et al. (2020) identified four categories of disease: COVID, SARS, ARDS, and both (COVID, ARDS). For these categories the modality of X-ray image samples are collected by John Hopkins University. In a patient record, there are 24 attributes, including name, offset, sex, age, finding, survival, intubation, need for supplemental oxygen, extubation, oxygen saturation, number of neutrophils, lymphocytes, view, mode, date, location, folder, filename, DOI, URL, License, Clinical notes, and other notes. On the basis of the textual attribute "Clinical notes," 40 features are determined. The proposed model makes use of Logistic Regression (LR) and Multinomial Nave Bayes (MNB). It has been tested on 212 patients. It is possible to combat this epidemic by analyzing the clinical reports and making appropriate recommendations.

Banik et al. (2021) have attempted to estimate the likelihood of a person contracting COVID-19. The classifiers like Linear Regression, Linear SVM, and Deep Learning are used for experimentation. As a result of these algorithms, an accurate model is developed. This model can be used to predict the probability that a patient is infected based on his or her clinical symptoms. This model considers several factors equally relevant. However, some characteristics are more pertinent to identifying COVID-19 cases [11]. Adhikari, M., and Agrawal, S. et al. (2001) presented an auto-regressive integrated moving average to deduce the k^{th} prediction of the experimental health signs of a patient, and Akaike Information Criteria are introduced to choose the best-fit prediction algorithm. For finding the positive and negative patients, the features are extracted, and the patterns are learned by the K-Means algorithm. Dataset has been generated synthetically by health symptoms of 400 patients. This dataset is used to measure the value of the proposed model and compared with standard methods [12]. Iwendi et al. (2020) used the AdaBoost algorithm to optimize the Random Forest model. COVID-19 can predict the severity of a patient's case based on geographic location, travel, health and demographic data. The dataset contains date, string, and numerical data types. The dataset is classified using a boosted random forest technique. This model has an accuracy rating of 94% , and F1 Score of 0.86. Data analysis indicates that it correlates positively with mortality. The number of deaths ranged from 20 to 70 years of age [13].

Proposed Methodology

Based on a Dynamic Scale Factor (DSF) based Elephant Herding Optimization (DSEHO), Covid 19 prediction is carried out integrating Weighted Support Vector Machines (WSVM) derived from the DSF. During this process, inputs are preprocessed, dimensions are reduced, features are selected, and classification is performed. The flow diagram is shown in Figure 1.

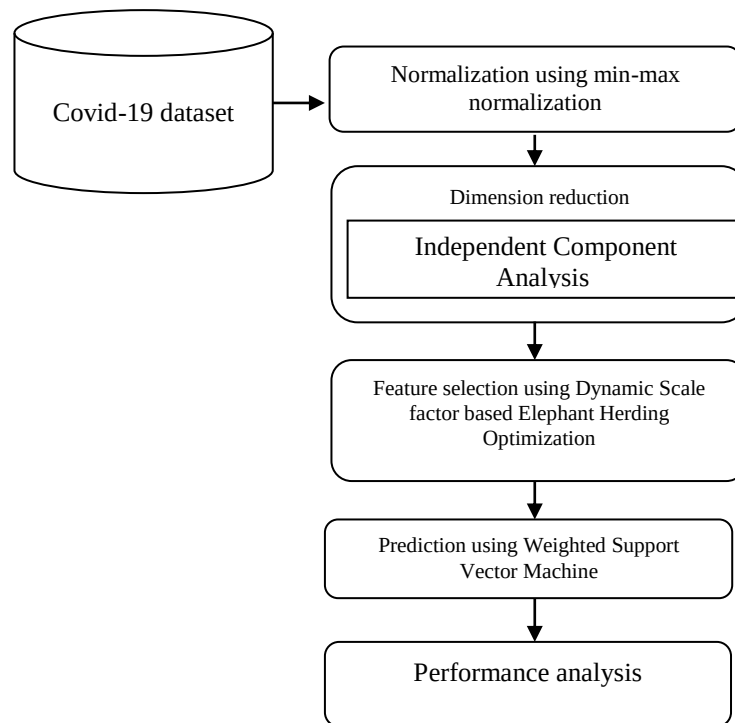


Figure 1. Flow diagram of the proposed work

Input

The dataset can be viewed at <http://www.kaggle.com/iamhungundji/covid19-symptoms-checker>. Each variable in the dataset motivation be related to the presence or absence of Corona virus disease. The variables include:

- Country: The following is a list of the countries the individual has visited.
- Age: According to the WHO Age Group Standard, each individual is classified based on their age
- Symptoms: There are several symptoms associated with COVID-19, including fever, fatigue, difficulty breathing, dry cough, and sore throat.
- Any other symptoms: Other symptoms may include pain, nasal congestion, a runny nose, diarrhoea, etc.
- Severity: In terms of severity, there are three levels: mild, moderate, and severe
- Contact: COVID-19 patients have been contacted by the individual responsible for contacting them

Data normalization

Min-max normalization is used for data normalization of the input dataset. An original the dataset is linearly altered though min-max normalization. This alters the values within a limited range of values. For mapping a v value, of an attribute A from range $[min_A, max_A]$ to a new range $[new_min_A, new_max_A]$, the computation is given by,

$$v' = \frac{v - x_{min}}{x_{max} - x_{min}} (new_x_{max} - new_x_{min}) + new_x_{min} \quad (1)$$

Where, v' - new value in the required range, x_{min} and x_{max} - minimum and maximum value of variables, new_x_{max} and new_x_{min} - new range of attributes

Algorithm 1: Min-max normalization algorithm

Input: COVID Dataset

Output: Normalized dataset

- Start
- Fix the maximum and the minimum bound
- Calculate the x_{min} and x_{max} values for samples
- 4: For each feature
- Calculate the values of v' by equation (1)
- End for
- End

Dimension reduction

Independent Component Analysis (ICA) is used to reduce the dimensions and it needs the coefficients of a linear expansion of the dataset vectors toward be mutually independent. In order to determine the ICA expansion, higher-order statistics are needed for dimensionality reduction. It was developed to distinguish independent sources as of their linear mixtures. Considering an $n \times p$ data matrix X (input dataset), whose rows $\mathbf{r}_i$ ($i=1, \dots, n$) match toward observational features and whose columns $\mathbf{c}_j$ ($j=1, \dots, p$) are in the ICA model of X , the individuals match toward the corresponding features,

$$X = AS \quad (2)$$

S is a source matrix whose rows are as statistically independent as possible, so that A is a mixing matrix with $n \times n$ and S is a source matrix with $n \times p$. These independent components are known as a linear mixture of independent components. The statistical independence among features be able to be calculated via mutual information $I = \sum_k H(s_k) - H(S)$ where, $H(s_k)$ is the marginal entropy of the feature s_k and $H(S)$ is the joint entropy. Finding the most appropriate linear combinations of observational features allows us to estimate the independent components since the mixing matrix can be inverted as follows:

$$U = S = A^{-1}X = WX \quad (3)$$

To reduce dimensions, this work the ICA algorithm is utilized. Contrast Function approximates the mutual information in this algorithm:

$$J(s_k) = (E\{G(s_k)\} - E(G(v)))^2 \quad (4)$$

Arbitrary non-quadratic function is denoted as G , normally distributed variable is denoted as v . Non-orthogonal basis is used in ICA to considering higher-order

dependencies in the data. ICA, therefore, is more sensitive to second-order relationships than PCA, which is based solely on first-order relationships. Conventionally, the independent components of the ICA model are typically scaled to unit deviations, while the signs and order of their components are arbitrary.

Feature selection using Dynamic Scale factor based Elephant Herding Optimization (DSEHO)

Using Dynamic Scale Factor-Based Elephant Herding Optimization (DSEHO), optimal features are determined after dimension reduction. Different optimization techniques are solved using the swarm-based search method, EHO. Elephants herd in groups, which played a critical role in the design of the EHO algorithm. Female elephants lead their clans. When an elephant reaches adulthood, she separates from her herd. There are therefore two operators within the elephant group, namely a separating operator and a clan updating operator. In order to solve different types of global optimization problems, a few general guidelines should be followed. These general guidelines are outlined next.

- Different elephant clans have different numbers of elephants.
- Every generation, a predetermined amount of male elephants will leave their families and live alone absent from the main herd.
- The matriarch of the clan is the one who leads the elephants.

Clan-updating Operator

Depending on their usual habits, elephant matriarchs lead their clans. Elephants' new positions are determined by their matriarchal influence. In the proposed study, the effects of matriarchal influence are examined using the dataset as input. Featured in the dataset are elephants. From the classification accuracy, a fitness function is derived. The feature j in the dataset can be calculated using Equation (5).

$$x_{new,ci,j} = x_{ci,j} + \alpha \times (x_{best,ci} - x_{ci,j}) \times r \quad (5)$$

Where, $x_{new,ci,j}$ and $x_{ci,j}$ shows the new location and old location for feature j in dataset c_i , correspondingly. $x_{best,ci}$ is matriarch c_i which signifies the optimal feature from the dataset. The $\alpha \in [0,1]$ specifies a scale factor, $r \in [0,1]$. The updating scale factor α of the EHO process is able to be transformed by the summation of groups in a linear manner by equation (6),

$$\alpha = \frac{2 \times \left(\frac{1-t}{t_{max}}\right)^2}{\left(\frac{1-\mu \times t}{t_{max}}\right)^3} \quad (6)$$

Where, t -present iteration, $iter_{max}$ - a maximum number of iterations, μ - adjustment coefficient in the interval [15, 35]. The best elephant is able to be designed by Equation (7) used for every clan.

$$x_{new,ci,j} = \beta \times x_{center,ci} \quad (7)$$

Where $\beta \in [0, 1]$ signifies a factor and it may controls the influence of the $x_{center, ci}$ on $x_{new, ci, j}$. $x_{new, ci, j}$ is the innovative individual. $x_{center, ci}$ is the center feature of dataset ci. It is able to be considered by Equation (8) for the dth dimension.

$$x_{center, ci, d} = \frac{1}{n_{ci}} \times \sum_{j=1}^{n_{ci}} x_{ci, j, d} \quad (8)$$

Where $1 \leq d \leq D$ and n_{ci} represented as the amount of features in dataset ci. $x_{ci, j, d}$ is denoted as the d-th dimension of feature $x_{ci, j}$.

Separating Operator

Separating operators are used to solve optimization problems in which male elephants leave the family group. Equation (9), as shown, separates the individuals in each generation with the lowest fitness.

$$x_{worst, ci} = x_{min} + (x_{max} - x_{min} + 1) \times rand \quad (9)$$

Where, x_{max} is represented as the upper bound of the feature and x_{min} is defined as lower bound of the feature. $x_{worst, ci}$ is defined as the worst feature in dataset ci. Stochastic distribution (Rand [0, 1]) is generated amongst 0 and 1.

Algorithm 2: Improved Scale Factor based Elephant Herding Optimization (ISFEHO)

Input: Number of features in the dataset

Output: Selected features

Step 1: Begin

Step 2: Initialization. Initialize the features in the dataset, set the initialize iterations $G = 1$; set maximum generation MaxGen.

Step 3: While stopping criterion is not met do

Step 4: Sort the population according to accuracy of individuals.

Step 5: For dataset ci do

Step 6: For each feature j in the dataset ci do

Step 7: Generate $x_{new, ci, j}$ and update $x_{ci, j}$ by Equation (5).

Step 8: If $x_{ci, j} = x_{best, ci}$ then

Step 9: Generate $x_{new, ci, j}$ and update $x_{ci, j}$ by Equation (7).

Step 10: End if

Step 11: End for

Step 12: End for

Step 13: For dataset ci do

Step 14: Replace the worst feature in the ci by Equation (9).

Step 15: End for

Step 16: Evaluate each features according to its position.

Step 17: $T = T + 1$.

Step 18: End while

Step 19: End.

Weighted Support Vector Machine (WSVM)

A weighted support vector machine (WSVM) is used to predict Covid-19. WSVM is a one of the machine learning algorithm with the purpose of classifies patterns in known features. Covid-19 is used to classify data according to the selected features. For classification, there are two classes of training features $(x_1, y_1), \dots, (x_n, y_n)$, where, $x_i \in R^n$ which signifies the i^{th} training features and $y_i \in \{-1, +1\}$ which characterizes the consistent class label. In order to achieve separation, one needs to discover a hyperplane $w \cdot x + b = 0$. In this weight vector ω is a usual vector of hyperplane, b is deviation, x is training features. The group formula is,

$$(w \cdot x) + b \geq \text{if } y_i = +1 \quad (10)$$

$$(w \cdot x) + b \leq \text{if } y_i = -1 \quad (11)$$

To get the best hyper plane, $\|w\|^2$ should be reduced under the constraint $y_i(w \cdot x_i + b) \geq 1, i = 1, 2, \dots, n$. With the introduction of positive slack variables ξ_i , the optimization problem is extended to include a nonlinear decision surface.

$$\min \frac{1}{2} \|w\|^2 \quad (12)$$

Subject to $y_i(x_i \cdot w + b) \geq 1$

WSVM assigns each sample a varied weight by its relative importance in the class, so that varied instances contribute in a different way toward the learning of the decision making. The WSVM with W_i as individual weights and form a new optimization issue is specified as

$$\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N W_i \xi_i \quad (13)$$

s.t $y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i=1, 2, \dots, n$

In equation (13), C is a penalty parameter which manages the balance among margin maximization and error reduction.

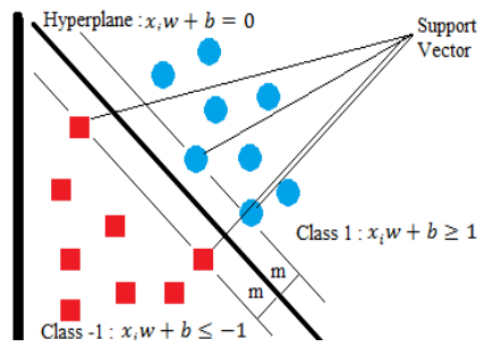


Figure 2. Illustration of classification in SVM

Experimental Results

The performance of the classification methods will be experimented and evaluated via Java. Dataset is collected from “<https://www.kaggle.com/iamhungundji/covid19-symptoms-checker>”. The proposed DSEHO-WSVM classifier outperforms the other classifiers such as XGBoost and SVM with respect to precision, recall, f-measure, and accuracy.

Table 1
Performance comparison of classification methods vs. metrics

Metrics	Classifiers		
	XGBoost	SVM	DSEHO -WSVM
Precision	92.94	95.51	96.56
Recall	91.86	94.44	95.52
F-measure	92.40	94.97	96.34
Accuracy	87.00	91.00	93.00

Accuracy

Accuracy is the most significant performance evaluation metric with the purpose of appropriately classifying the data of events and it is basically a ratio of expected exact examination to the complete examinations.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (14)$$

Where, *TP* - True Positive, *FN* - False Negative, *FP* - False Positive, and *TN*- True Negative

Precision

Precision is represented as the significance of the results which is calculated via the ratio of positive examinations predicted appropriately to the true examinations predicted in total.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (15)$$

Recall

Recall is the ratio of appropriately predicted true examination to the all examination in real class - yes.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (16)$$

F-measure

F-measure is represented as the weighted harmonic mean of precision and recall. It is described as follows,

$$F\text{-measure} = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \quad (17)$$

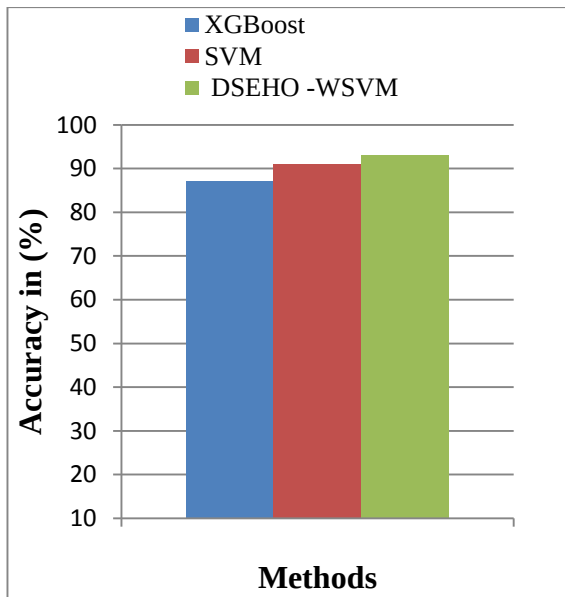


Figure 3. Accuracy comparison

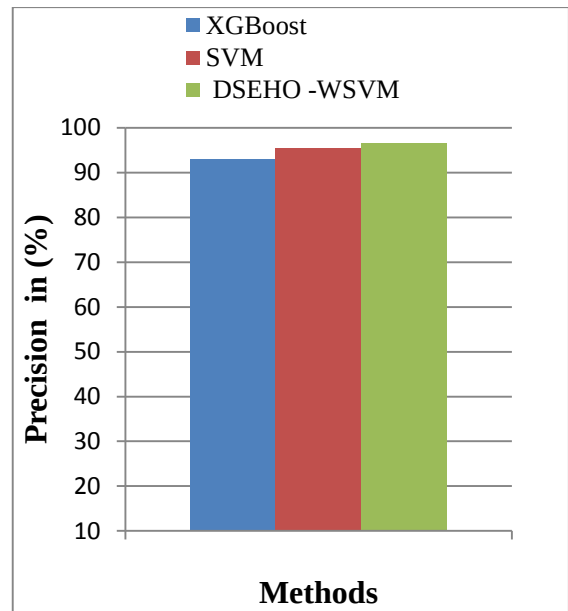


Figure 4. Precision comparison

Figure 3 and 4 represents a performance comparison of the proposed DSEHO-WSVM, XGBoost, and SVM. In these figures, classifiers used for comparison are taken in x-axis and metric results are discussed in y-axis. DSEHO algorithm is used for best feature selection from the dataset which increases the classification results of the classifiers. It increases the accuracy and precision results. Results demonstrated that the developed DSEHO-WSVM classifier achieves 93% of accuracy and 96.56% of precision which is increased than the existing classifiers such as XGBoost and SVM methods.

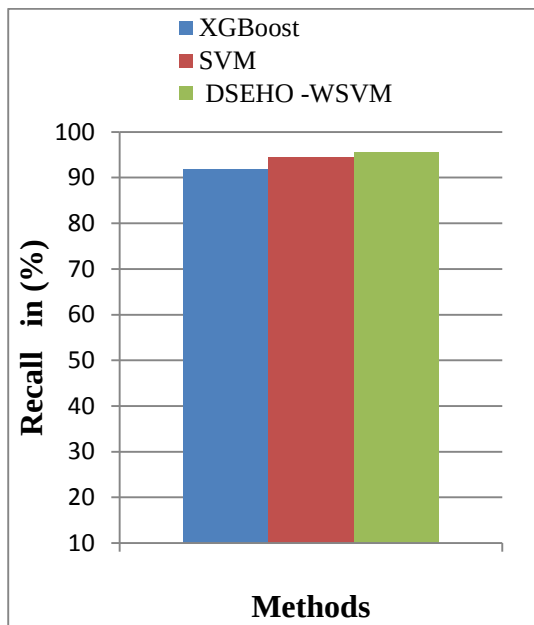


Figure 5. Recall comparison

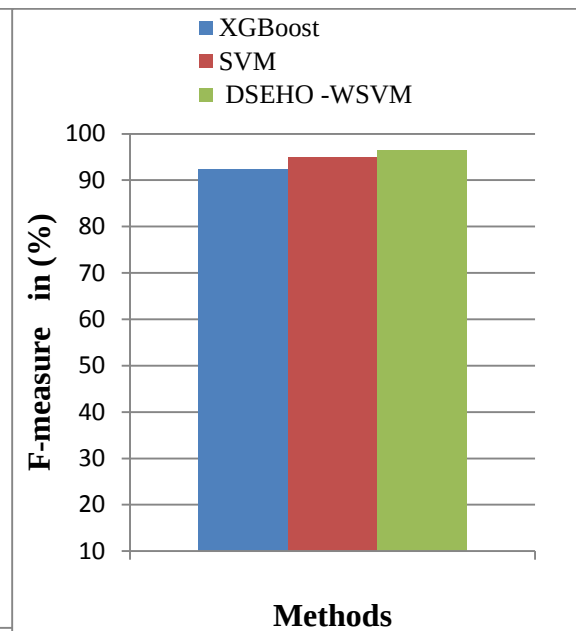


Figure 6. F-measure comparison

Figure 5 and Figure 6 shows the recall and f-measure comparison of the proposed DSEHO- WSVM, XGBoost, and SVM. In these figures, classifiers used for comparison are taken in x-axis and metric results are discussed in y-axis. DSEHO algorithm is used for best feature selection from the dataset which increases the classification results of the classifiers. It increases the accuracy and precision results. Results demonstrated that the proposed DSEHO- WSVM system achieves 95.52% of accuracy and 96.34% of f-measure which is increased than the existing classifiers such as XGBoost and SVM methods.

Conclusion

In the proposed system, the proposed the use of Dynamic Scale Factor based Elephant Herding Optimization (DSEHO) with Weighted Support Vector Machines to predict COVID-19 coefficients. Min-max normalization is first performed on the input dataset. Elephant Herding Optimization (DSEHO) is an algorithm designed to improve classification accuracy by using Dynamic Scale factors. Finally, weighted support vector machines (WSVM) are used to predict covid-19. The results of the classification methods are evaluated via the metrics like precision, recall, f-measure, and accuracy. Thus the proposed system outperforms than the previous methods.

References

1. Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., ... & Ferguson, N. M. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6), 669-677.

2. Azarpazhooh, M. R., Morovatdar, N., Avan, A., Phan, T. G., Divani, A. A., Yassi, N., ... & Di Napoli, M. (2020). COVID-19 pandemic and burden of non-communicable diseases: an ecological study on data of 185 countries. *Journal of Stroke and Cerebrovascular Diseases*, 29(9), 105089.
3. Mohamadou, Y., Halidou, A., & Kapen, P. T. (2020). A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. *Applied Intelligence*, 50(11), 3913-3925.
4. Gothai, E., Thamilselvan, R., Rajalaxmi, R. R., Sadana, R. M., Ragavi, A., & Sakthivel, R. (2021). Prediction of covid-19 growth and trend using machine learning approach. *Materials Today: Proceedings*.
5. Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., ... & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature medicine*, 26(7), pp.1037-1040.
6. Friston KJ, Parr T, Zeidman P, Razi A, Flandin G, Daunizeau J, et al. Dynamic causal modelling of COVID-19. *arXiv preprint arXiv:200404463*. 2020.
7. Wynants, B. Van Calster, M. M. Bonten, G. S. Collins, T. P. Debray, M. De Vos, et al., "Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection," medRxiv, 2020.
8. G. Pandey, P. Chaudhary, R. Gupta, and S. Pal, "SEIR and Regression Model based COVID-19 outbreak predictions in India," arXiv preprint arXiv:2004.00958, 2020.
9. Guhathakurata, S., Kundu, S., Chakraborty, A., & Banerjee, J. S. (2021). A novel approach to predict COVID-19 using support vector machine. In *Data Science for COVID-19* (pp. 351-364). Academic Press.
10. Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*, 12(3), 731-739.
11. Banik, S., Banik, S., Ghosh, A., & Mukherjee, A. (2021). Probabilistic estimation of COVID-19 using patient's symptoms. In *Data Driven Approach Towards Disruptive Technologies* (pp. 369-378). Springer, Singapore.
12. Nandy, S., & Adhikari, M. (2021). Intelligent Health Monitoring System for Detection of Symptomatic/Asymptomatic COVID-19 Patient. *IEEE Sensors Journal*, 21(18), 20504-20511.
13. Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., ... & Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in public health*, 8, 357.