Classification of COVID-19 by using supervised optimized machine learning technique

Dilip Kumar Sharma [a], Muthukumar Subramanian [b], Pacha. Malyadri [c], Bojja Suryanarayana Reddy [d], Mukta Sharma [e],*, Madiha Tahreem [f]

[a] Department of Mathematics, Jaypee University of Engineering and Technology, Guna, M.P., India
[b] Nagarjuna College of Engineering & Technology, Bangalore, India
[c] Centre for Economic and Social Studies, Begumpet, Hyderabad, Telangana, India
[d] Jaagruthi Degree and PG College, Mahatma Gandhi University, Bhongir, Yadadribhongir, Telangana, India
[e] School of Life Sciences and Technology, IIMT University, Meerut.
[f] Zakir Hussain College of Engineering & Technology, Aligarh Muslim University, Aligarh, U.P., India

ARTICLE INFO

ABSTRACT

In recent two years, covid-19 diseases is the most harmful diseases in entire world. This disease increase the high mortality rate in several developed countries. Earlier identification of covid-19 symptoms can avoid the over illness or death. However, there are several researchers are introduced different methodology to identification of diseases symptoms. But, identification and classification of covid-19 diseases is the difficult task for every researchers and doctors. In this modern world, machine learning techniques is useful for several medical applications. This study is more focused in applying machine learning classifier model as SVM for classification of diseases. By improve the classification accuracy of the classifier by using hyper parameter optimization technique as modified cuckoo search algorithm. High dimensional data have unrelated, misleading features, which maximize the search space size subsequent in struggle to process data further thus not contributing to the learning practise, So we used a hybrid feature selection technique as mRMR (Minimum Redundancy Maximum Relevance) algorithm. The experiment is conducted by using UCI machine learning repository dataset. The classifier is conducted to classify the two set of classes such as COVID-19, and normal cases. The proposed model performance is analysed by using different parametric metrics, which are explained in result section.

1. Introduction

COVID-19 is related to a range of other contagious diseases and the most communal symptoms such as fever and cough make accurate analysis create critical blocks for the health professionals. It is noted that the RT-PCR, is considered to be more accurare diagnostic test, are often used for over a week to be available in Brazil, based on reports, while immediate decisions on clinical treatment and preventive actions are needed in the meantime. On the other hand, new fast diagnostic tests, which are prone to certain precise problems, have recently increased their usage, and this could upsurge the risk of ineffective health resources [1,2].

Effective screening makes it possible to diagnose COVID-19 fast and efficiently and can reduce the cost to healthcare systems. Prediction models have been developed with a view to helping medic's worldwide triage of patients, in particular in light of limited health resources that incorporate various functions to evaluate infection risks [3]. These simulations use features as CT scans, clinical symptoms, laboratory testing and the integration. Most models have been based, however, on data obtained from hospitalized patients so SARS-CoV-2 screening of the general population is not effective (see Fig. 1).

The first group demonstrates that the majority of studies focussed on the prediction of COVID-19 infection with meteorological information by assessing relevant trips in each cluster. The sec-

* Corresponding author.
E-mail address: muktavats@yahoo.com (M. Sharma).

ond group shows that the rest of the research has emphasized the use of deep learning algorithms for various chest related CT scan and X-rays [4]. Although COVID-19 has been characterised by excellent sensitivity through CT and X ray images [5] these tests may, because of the radiation dose, high costs and a low number of devices available, sometimes be problematic to employ for patients' screening. Therefore, it remains a dare that needs to be curb the pandemic to distinguish between positive and negative COVID-19 cases. Effective screening makes it possible to diagnose COVID-19 quickly and efficiently and can reduce the load on health systems [6]. Prediction models that use multiple characteristics to evaluate the risk of infection in order to help medical workers throughout the globe trial patients with limited healthcare resources were developed (see Table 1).

2. Literature survey

Huang et al [7] has created a forecasting model through the application of SVM model which is then combined with RBF kernel so as to forecast the overall readmission rate related to cured pneumonia of the individuals who have been moved out of the hospital after treatment. This analysis enables in forecasting with the accuracy rate of more than 82% and can be considered as the effective tool in analysing the individuals with pneumonia.

Lee et al. [8] used the bayesian feature selection approach and subsequently employed Leukemia dataset to apply ANN, KNN and SVM classifications. Ye et al. [9] employed ULDA for the selec-

distinguish COVID-19 and non-COVID with a patient who is safely and bacterially contaminated.

To identify COVID viral infection, Sethy et al. [14] used the chest X-ray pictures to initially extract the deep functionality using CNN, based on retrained ImageNet and to categorize the layer last SVM.

3. Proposed methodology

In this section we discussed the proposed methodology and materials of this study. In following section we introduced the proposed methods schemes as mRMR for feature selection, modified cuckoo search algorithm for machine learning parameter optimization. Here we classify the cases as positive or negative by using optimized SVM classifier method.

4. Minimum redundancy and maximum relevance for feature selection

Distinct from the univariate model in order to choose from the different features without considering the overall redundancy on the chosen aspects, the MRMR function chooses the appropriate function which will enable in considering the most critical task of forecasting and is also poised in reducing the redundancy through the various features which has been selected.

The below algo 1 states the MRMR function related to the key feature selection for the chosen problems:

Algorithm 1: mRMR Algorithm

```
CHOSEN Δ Ø
FS Δ index list of DS
While FS! =  NULL Do
max eval Δ  − 1
idx Δ NULL
For i Δ 0 To length of FS Do
tmp flist Δ {CHOSEN ∪ FS [i]}
tmp_eval Δ mrmr (tmp_flist) //evaluate given feature subset by mRMR measure
IF tmp eval > max eval Then
max eval Δ tmp eval
idx Δ i
End If
End For
Append FS [idx] To CHOSEN
Delete FS [idx] From FS
End While
Return CHOSEN
```

tion of features, and gathered a higher level of classification compared with previous approaches. Selection and kernels based fuzzy classification methods were used in SVM-RFE by Cho et al. [10].

The diagnostic paradigm of COVID-20 based on clinical and radiological characteristics has been proposed and validated by Chen et al. [11]. Burian et al. [12] assessed clinical symptoms and the imaging characteristics of the need for ICU therapy. Two further studies have been undertaken on the basis of the blood tests to identify the positive COVID-19 cases.

Ref. [13] was provided as a multi-class, deep-recovery architecture of the neural network known as COVID-Net with 16,756 thoroughly scanned pictures of 13,645 patient chest radiation to

5. Hyper-parameter optimization

The efficient search of the space of the hyperparameters utilizing optimization techniques can identify ideal hyperparameters for models during the project process of the ML prototypes. The process of optimization for hyper-parameters comprises of four basic components: an estimator (a regression or a gradation) with their goal function, a search space (configuration area). Searching or optimization scheme used to locate hyper-parameter arrangements.

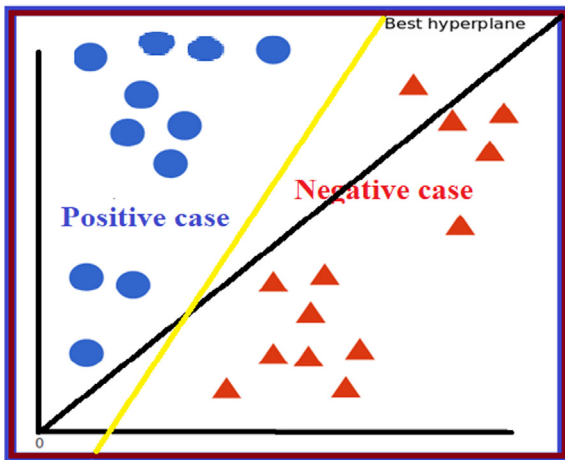The chief process of hyper-parameter arrangements is as follows:

Fig. 1. Represents a simple SVM for lung cancer.

1. Choose objective function and performance measurements;
2. Choose the parameters to be tuned, review their types and identify the best approach of optimization;
3. Provide the necessary workings for the ML model as baseline model using default setup of parameters or common values;
4. Initialize optimization with a wide search area as the viable hyper parameter domain defined by manual testing skills.
5. Narrow search area on the basis of the regions with well-testing and explore new search areas.
6. Return the most successful configuration of the hyperparameter as final solution

The optimum hyperparameters for the templates can be found during the design phase for model ML by successfully searching the space of the hyperparameters with optimization algorithms. The process of offering better parameter optimization is comprised of critical aspects which covers that the predictor possess certain objectives and outcomes, there needs a space which is used to apprehend the data and also optimise the combination, it also involves in application of various assessment function which will enable in comparing the presentation. The optimization process consists of four main components; [15]

5.1. Modified cuckoo search algorithm

This model intends to create an algorithm based on the cuckoo, which is mainly due to the brood parasites. These birds do not tend to build their own nest ad lays their eggs in host bird. The bird is more of a parasite of the best known aspect, the cuckoo tend to involve with the host birds, in case if the host finds that these are not their, they might throw it away and delete the home and create a new one. So, each aspect of the egg in the nest tend to state he overall solution. [16]

The overall results which are mainly received through these aspects on the current ones and there exist alteration of the same functionalities. Each of the nest tend to contain different eggs and hence CS enable in addressing the various issues and can be

applied in structural engineering for effective optimisation. It also enable in application of the reorganisation of expression, task aspects and global optimisation. [17]

1. Each of the bird lays one egg at a given time and also places them in the nest on a random basis.
2. The next generation best nest will be made available with good egg quality. [18]
3. The number and likelihood of pa = [0, 1] are determined by the sum of host nests accessible, and if a hostbird identifies the cuckoo egg, it can either toss it away or leave and create a new nest. In following algorithm for Modified Cuckoo Search Algorithm.

$Input : Population\ of\ nests\ x_i = (x_{i1}, \cdots, x_{iD})T$
$\quad for\ i = 1 \cdots Np, MAX\_FE.$
$Output : The\ best\ solution\ x_{best}\ and\ its$
$\quad corresponding\ value\ f_{min} = min(f(x)).$
$Step.1 : generate\ initial\ host\ nest\ locations();$
$Step.2 : eval = 0;$
$Step.3 : while\ termination\ condition\ not\ meet\ do$
$Step.4 : for\ i = 1\ to\ Np\ do$
$Step.5 : x_i = generate\_new\_solution(x_i);$
$Step.6 : f_i = evaluate\_the\_new\_solution(x_i);$
$Step.7 : eval = eval + 1;$
$Step.8 : j = [rand(0,1) * Np + 1];$
$Step.9 : if f_i < f_j then$
$Step.10 : x_j = x_i; f_j = f_i; //\ replace\ j - th\ solution$
$Step.11 : end\ if$
$Step.12 : if\ rand(0,1) < pa\ then$
$Step.13 : init\_nest(x_{worst});$
$Step.14 : end\ if$
$Step.15 : if\ f_i < f_{min}\ then$
$Step.16 : x_{best} = x_i; f_{min} = f_i; //save\ the\ local\ best\ sol.$
$Step.17 : end\ if$
$Step.18 : end\ for$
$Step.19 : end\ while$

6. Machine learning classification

Here we classify the dataset cases into two order such as positive or negative by using SVM classifier.

6.1. Support vector machine

A SVM is a monitored learning algorithm which can be utilized for problems like classification and regression. SVM techniques are based on the mapping notion of low-dimensional data points and are linearly separated by a high-dimensional space. A hyperplane

Table 1
Hyper parameter of different machine learning models.

S.no	Related algorithm	Critical Parameters	Optional	HPO methods
1	SVM	C, kernel, Sigma, epsilon	C, kernel, Sigma, epsilon	Modified Cuckoo search Algorithm

Dilip Kumar Sharma, M. Subramanian, Pacha. Malyadri et al.

Materials Today: Proceedings 56 (2022) 2058–2062

is then formed as a classification limit for the partition data points. Algorithm 3 demonstrated as Support Vector Based Classification.

Algorithm 3. Support Vector Based Classification (SVM)

Input : *TrainingSetasδ* :
   ThresholdOutput : *XR* : *XRCX*
   |XR| << |X| Begin Train a decision tree T;
   //XR Begins empty XRΔNULL
   ForeachleafLiofTdo
   foreachoppositeclassneighborLjdo
   ifentropyofLiislowthen
   //Selectclosestexamples
   UseLiLjtobuildX+;
   Computeω
   AddxjϵLjtoXR
   endfor
   else
   //AddalltheelementsinLjtoXR
   XRΔXR ∪ Lj;
   endifendforreturnXREnd

THE SVM objective function as.

$$\arg \underbrace{\min}_{W}\left(\frac{1}{n}\sum_{i=1}^{n}\max\{0; 1 - y_i f(x_i)\} + Cw^T w\right)$$

where w is a vector for standardization; C is an error term penalty parameter that is important in all the SVM models. The kernel function f(x) can be selected from a variety of kernel types in SVM models to measure the likeness among two x_i and x_j data points. The kernel type would therefore be a significant hyperparameter. Common SVM kernel types include linear kernels, the RBF, multi- kernels and sigmoid kernels. You can describe the various kernel functions as follows

6.2. SVM kernel functions

6.2.1. Radial basis function kernel

The RBF kernel agrees to an infinite dimensional feature space f. The RBF function is given by,

$$k(x_i, x_j) = \left(\frac{\exp(-\gamma)\|x_{i,-}x_j\|^2}{2\sigma^2}\right)$$

where ϕ is a mapping function, when it is identity, no mapping (i.e., cannot actually write down the vector ϕ x).

6.2.2. Linear kernel

It is one of the simplest kernel function that is represented by,

$$k(x_i, x_j) = x_{,i}^T x_j$$

Which is the inner product in addition to an elective constant c. The Kernel algorithms which use linear kernel are typically the same for their non-kernel counterparts.

6.2.3. Quadratic kernel

The quadratic kernel function is given by,

$$k(x_i, x_j) = (x_{,i}^T x_j)^2$$

6.2.4. Polynomial kernel

The Polynomial kernel is one of the non-stationary kernels which are appropriate for all the training data that are normalized. The parameters which are editable include α - the slope, d- polynomial degree and c – the constant term.

$$k(x_i x_j) = (ax_i^T x_j + c)^2$$

Since the vectors in the Gaussian radial basis function are mapped nonlinearly to a very high dimensional feature space, it is found to be very useful when compared to other kernel functions.

A few distinct hyper functions must be tweaked after choosing a kernel type, as illustrated in the kernel functional equations. The coefficient α, designated in sklearn as 'gamma' means a conditional hyperparameter for the hyperparameter 'kernel type' when it is set to Polynomial, RBF or Sigmoid; r, described in Sklearn as 'coef0.'.. In addition, an extra conditional hyper-parameter d is available to the polynomial kernel denoting the 'degree' of the polynomial func- tion. There is additional hyper parameter, 'epsilon,' for support vec- tor regression models that shows the distance error from its loss function.

7. Results and discussion

In this study, SVM classifier with various kernel functions is used to classify the different types of covid and non-covid images. It depends on the type of feature selection technique chosen and the hyper parameter optimization. This dataset consists of toally 1000 images, which 300 images due to pandemic images and 700 images negative aspects. We train and test the model by split the entire dataset into 70:30 ratio.

7.1. Performance measures

The overall output is used to estimate various metrics of the classifier. True positive (TP) is a COVID-19 positive image correctly identified as COVID-19 positive. The implicationof the negative is related to the pandemic image correctly identified as negative COVID-19. False positive (FP) is a negative COVID-19 image incor- rectly identified as COVID-19 positive. False negative (FN) is an COVID-19 positive image incorrectly identified as negative COVID-19. All the classifier quality parameters are measured by TP, TN, FP and FN. The classifier performance is evaluated by the following performance measures

Specificity: It refers to the ability to find out negative results of the classifier i.e percentage of the abnormal images incorrectly identified as normal.

$$\text{Specificity} = TN/(TN + FP) \tag{1}$$

Sensitivity: It refers to the ability to find out positive results of the classifier that is percentage of the abnormal images correctly identified as abnormal.

$$\text{Sensitivity} = TP/(TP + FN) \tag{2}$$

Precision: It is represent as the positive predictive value is the probability that a positive prediction is correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

F_score: It is a harmonic mean of precision and specificity, the F-measure.

$$\text{F\_score} = 2 * \frac{(\text{Precision} * \text{Sensitivity})}{\text{Precision} + \text{Sensitivity}} \tag{4}$$

Dilip Kumar Sharma, M. Subramanian, Pacha. Malyadri et al.

Materials Today: Proceedings 56 (2022) 2058–2062

Table 2
Performance analysis of SVM classifier with different technique.

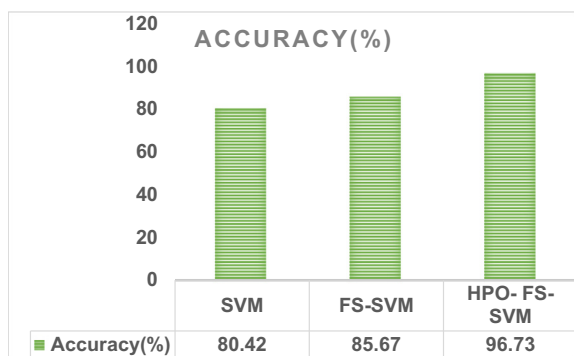| Methods	Precession (%)	Sensitivity (%)	Specificity (%)	F_score (%)
SVM	80.42	85.52	90.31	94.45
FS-SVM	85.67	87.59	94.21	95.11
HPO- FS-SVM	96.73	97.15	96.15	98.50

Fig. 2. Accuracy performance analysis of proposed method.

Accuracy: It is calculated as the proportion of the correctly predicted total number of images as correctly.

$$Accuracy = \frac{Sensitivity + specificity}{2} \qquad (5)$$

In Table 2 represent that the performance analysis of diiferent order of proposed method. Initially, we calculate the performance by using normal SVM, which achieve the Precession of 80. 42(%), F_score of 94.45%. Another feature selection technique is combined with SVM, which achieve preceison of 85.67% and F-score of 95.11%. finally, hyperparameter tuning of SVM achieve the better classification results as precision value of 96.73%, sensitivity of 97.15%, specificity of 96.15% and F0score of 98.50%. by this comparision, HPO- FS-SVM scheme achieved the better performance than other scheme.

In Fig. 2 signified the graphical representation of accuracy analysis of different scheme. Initially, we calculate the performance by using normal SVM, which achieve the accuracy of 80.42%. Another feature selection technique is combined with SVM, which achieve the accuracy of 85.67%. Finally, hyper parameter tuning of SVM with feature selection scheme achieve the better classification accuracy of 96.73%, By this comparison, HPO- FS-SVM scheme achieved the better performance than other scheme.

8. Conclusion

In this study, we projected a technique for COVID-19 cases diagnosis on different positive and negative case images. We evaluated this study by used the hyper parameter optimization of machine learning classifier as SVM. To optimize the SVM classifier parameter by using Modified cuckoo search algorithm to achieve the better classification results. A hybrid feature selection technique as mRMR algorithm was used for feature selection. This framework has been validated against publicly available covid-19 datasets and further optimized using modified cuckoo search algorithm. For experimental analysis two sets of data are used for training and testing. Comparability, sensitivity, specificity and precision assessment metrics of the least amount of characteristics were attained with the proposed method. By identifying the most significant traits, the methodology proposed achieved both high performance and resource use. Other medical and other important applications can be included in our future work.

CRediT authorship contribution statement

Dilip Kumar Sharma: Investigation, Writing – original draft. **Muthukumar Subramanian:** Conceptualization, Writing – review & editing, Supervision. **Pacha. Malyadri:** Formal analysis, Data curation. **Bojja Suryanarayana Reddy:** Conceptualization. **Mukta Sharma:** Writing – review & editing. **Madiha Tahreem:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.F. de Moraes Batista, J.L. Miraglia, T.H. Donato, A.D. Chiavegatto Filho, COVID-19 diagnosis prediction in emergency care patients: a machine learning approach, MedRxiv (2020).

[2] M. Kukar, G. Gunčar, T. Vovko, S. Podnar, P. Černelč, M. Brvar, M. Zalaznik, M. Notar, S. Moškon, M. Notar, COVID-19 diagnosis by routine blood tests using machine learning, Sci. Rep. 11 (1) (2021) 1–9.

[3] A.Z. Khuzani, M. Heidari, S.A. Shariati, COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images, Sci. Rep. 11 (1) (2021) 1–6.

[4] I. Arpaci, S. Huang, M. Al-Emran, M.N. Al-Kabi, M. Peng, Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms, Multimedia Tools Appl. 80 (8) (2021) 11943–11957.

[5] Y. Zoabi, N. Shomron, COVID-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach, MedRxiv (2020).

[6] H.B. Syeda, M. Syed, K.W. Sexton, S. Syed, S. Begum, F. Syed, F. Prior, F. Yu Jr, Role of machine learning techniques to tackle the COVID-19 crisis: Systematic review, JMIR Med. Inform. 9 (1) (2021) e23811.

[7] J.S. Huang, Y.F. Chen, J.C. Hsu, Design of a clinical decision support model for predicting pneumonia readmission, in: International Symposium on Computer Consumer and Control, 2014, pp. 1179–1182.

[8] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: A bayesian variable selection approach, Bioinformatics 19 (1) (2003) 90–97.

[9] J. Ye, T. Li, T. Xiong, R. Janardan, Using uncorrelated discriminant analysis for tissue classification with gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 1 (2004) 181.

[10] J.-H. Cho, D. Lee, J.H. Park, I.-B. Lee, Gene selection and classification from microarray data using kernel machine, FEBS Lett. 571 (2004) 93.

[11] X. Chen, Y. Tang, Y. Mo, S. Li, D. Lin, Z. Yang, Z. Yang, H. Sun, J. Qiu, Y. Liao, J. Xiao, X. Chen, X. Wu, R. Wu, Z. Dai, A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study, Eur. Radiol. 30 (9) (2020) 4893–4902.

[12] E. Burian et al., Intensive care risk estimation in COVID-19 pneumonia based on clinical and imaging parameters: experiences from the munich cohort, J. Clin. Med. 9 (5) (2020) 1514.

[13] L. Wang, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. arXiv preprintaarXiv preprintarXiv preprint arXiv:2003.09871 (2020).

[14] P.K. Sethy, S.K. Behera, Detection of coronavirus disease (covid-19) based on deep features (2020).

[15] J. Ankit, A.K. Yadav, Y. Shrivastava, Modelling and optimization of different quality characteristics in electric discharge drilling of titanium alloy sheet, Mater. Today Proc. 21 (2019) 1680–1684.

[16] A. Jain, A. Kumar Pandey, Modeling and optimizing of different quality characteristics in electrical discharge drilling of titanium alloy (Grade-5) sheet, Mater. Today Proc. 18 (2019) 182–191.

[17] J. Ankit, A.K. Pandey, Multiple quality optimizations in electrical discharge drilling of mild steel sheet, Mater. Today Proc. 8 (2019) 7252–7261.

[18] V. Panwar, D.K. Sharma, K.V.P. Kumar, A. Jain, C. Thakar, Experimental Investigations And Optimization Of Surface Roughness In Turning Of EN 36 Alloy Steel Using Response Surface Methodology And Genetic Algorithm, Mater. Today: Proc. (2021), https://doi.org/10.1016/J.Matpr.2021.03.642.