Volume 66, Issue 2, 2022
Journal of Scientific Research
of
The Banaras Hindu University

An Optimal Random Forest Classifier for Diagnosing Covid-19 from X-ray and CT-scan Images

Sunil Kumar[1][0000-0003-1758-5204], Saroj Ratnoo [2][0000-0002-6083-4109]

[1]Guru Jambheshwar University of Science and Technology, Haryana, India,
skvermacse@gmail.com

[2] Guru Jambheshwar University of Science and Technology, Haryana, India ratnoo.saroj@gmail.com

Abstract: The unprecedented rate of spread and gravity of Covid-19 infection has played havoc worldwide. The spread of Covid-19 has pushed the nations' medical infrastructures to their limits, yet no one could contain the infection and loss of lives. Therefore, accurate and efficient diagnosis of the infection is of central importance to battle out the Covid-19 disease. RT-PCR, the customary diagnostic tool for Covid-19, is a time-consuming process. The RT-PCR results may take two to three days. In the wake of the new Covid-19 variants, the RT-PCR test may also result in false negative cases. Machine learning algorithms have been successful for auto diagnosing Covid-19 from epidemiological and medical image data. Applying a machine learning algorithm for Covid-19 detection from X-rays and CT scans images combined with RT-PCR results can augment the rate of Covid-19 diagnosis. Therefore, designing machine learning algorithms for the early Covid-19 diagnosis is the need of an hour. In this paper, we present Random Forest (RF) classifier for Covid-19 diagnosis. We use the resampling techniques to resolve the class imbalance in the data. Combining the RF classifier with its fine-tuned hyperparameters and resampling techniques to address the class imbalance yield promising results. Moreover, a Genetic Algorithm is employed to find an optimal hyperparameter configuration for the RF classifier. The statistical result of the approach indicates that accuracy, geometric mean and f-measure are 0.94, 0.93 and 0.92 respectively. The suggested method achieves higher sensitivity and comparable performance to deep and transfer learning approaches which are relatively computationally expensive but less comprehensible.

Keywords: Addressing Class imbalance, Covid-19 diagnosis from medical images, Hyperparameter tuning, Random Forest classifier.

1 Introduction

COVID-19 is a particular type of virus from the family of viruses known as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory System (MERS) [14, 48]. After SARS and MERS, COVID-19 is the third zoonotic disease which was declared a global pandemic on 11th March 2020 by WHO. As of 18th June 2021, the disease has infected 176,945,596 people globally, out of which 3,836,828 have lost their lives. The second wave of Covid-19 has recently taken its toll in India. There have been 29,700,313 confirmed cases in India alone, and the death toll has risen to 381903. Be it developed or developing nations, the mortality rate because of Covid-19 infections has been extremely high compared to the other typical influenza diseases. We have witnessed the healthcare systems struggling to contain the disease and save lives since the appearance of this virus. The worldwide experience shows that once Covid-19 starts spreading, it takes very little time to quash the health/medical care systems. At the same time, healthcare systems and medical experts are gaining experience to contain the disasters of Covid-19 disease. By now, a lot of epidemiological and medical imaging data of Covid-19 patients are accessible. Advanced machine learning algorithms can play a significant role in building predictive models for the prognosis and diagnosis of the Covid-19 patients to assist medical practitioners and policymakers [40, 65].

Machine Learning can learn valuable relationships and patterns from volumes of data without human intervention. Machine learning algorithms like Decision Trees (DT), Naive Bayesian (NB) and Bayesian Belief Networks (BBN), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Neural Networks (NNs) and Random Forests (RF), have been extensively used as the classifiers for Covid-19 diagnosis [27, 39, 47, 63]. In addition, deep and transfer learning techniques have also been employed for automatic Covid-19 disease diagnosis [1, 7, 8, 30, 33, 43, 56]. Deep learning approaches auto-extract features and achieves high accuracy. However, these approaches are computationally expensive, need a considerable amount of training examples, and are not suitable for human analysis and insight. Therefore, deep learning should be applied if the simple machine learning techniques do not achieve the required level of accuracy.

Out of the classical classification approaches, the decision trees are popular classifiers due to their reasonable accuracy and comprehensibility. However, decision tree construction adopts a greedy approach during the training phase and is prone to making sub-optimal predictions. To address this shortcoming, the data mining community has developed ensemble classifiers. An ensemble classifier is a set of many individual classifiers. In an ensemble approach, the individual decisions of the classifiers are aggregated in some way to classify the new examples. For instance, Random Forest (RF) is an example of an ensemble classifier. An RF consists of hundreds of decision trees built from different subsets of training

DOI: 10.37398/JSR.2022.660225

instances and predictors. An RF creates trees using a resampling process called bagging, and each node of a tree in the forest is split based on a randomly selected subset of all attributes. One of the methods for making predictions is majority voting. Since an ensemble is built by applying the local search algorithms from many starting points, the ensemble classifiers better approximate the proper unknown function between the dependent attributes and the class attribute. Therefore ensemble classifiers such as RF are more accurate than any individual classifier [25]. A few applications of ensemble classifiers exist for diagnosis Covid-19 from medical images [2, 31, 34, 59, 62, 66].

X-ray and CT scan images are widely used for primary pneumonia disease diagnosis. The most reliable test for Covid-19 diagnosis is Reverse Transcription Polymerase Chain Reaction (RT-PCR). But, RT-PCR results may take time (2 to 3 days) when an enormously large crowd is tested during a mounting Covid-19 wave. The delayed test results delay the treatment of presumptive patients. Further, RT-PCR can have high false negative cases in the presence of new variants since Chest X-rays present visual indices associated with Covid-19. Many studies show that chest X-ray images also provide a rapid and sensitive Covid-19 induced pneumonia diagnosis [34, 36, 38, 55]. Though expensive compared to X-ray images, CT scan imaging can also be used as another effective tool for COVID-19 disease diagnosis [8]. The classifiers trained on X-ray and CT scan images can be supportive in pre-screening the patients for faster decision-making before the availability of RT-PCR results. The combination of RT-PCR results and application of machine learning algorithms on medical images can speed up the proficiency of Covid-19 diagnosis, and it can contain the damage of letting go of the false negative cases untreated.

The medical image datasets may suffer from the problem of class imbalance that needs to be addressed. Class imbalance refers to the skewed class distribution in the datasets, i.e., In a medical dataset, only a few patients may test positive compared to those who test negative. The skewed class distribution influences the performance of a classifier negatively, and the classifier tends to favor the majority class (absence of disease). Therefore, such classifiers have a high false-negative rate and predict diseased people as the non-diseased ones. The false-negative predictions may prove to be very costly for a fatal and infectious disease like Covid-19 regarding the spread of disease and mortality. The class imbalance is usually addressed through resampling techniques which restore the balance between the instances of different classes. Some of the standard methods for resampling are undersampling, oversampling, and Synthetic Oversampling Minority Technique (SMOTE), Majority Weighted Minority Oversampling Technique (MWMOTE), and Random walk Oversampling (RWO) [10, 13, 70].

We have applied an RF algorithm for Covid-19 diagnosis using epidemiology labeled images of X-ray and CT scan in the current research. We have used the Grey Level Co-occurrence Matrix (GLCM) technique for feature extraction. GLCM features extraction method uses a second-order statistics approach to determine the overall average for the correlation degree between two pixels in different expects such as contrast, correlation, energy, homogeneity etc. GLCM approach help to reduce the compression time of the image to a great extent for converting RGB to Grey levels and also produce the fewer essential features of images [42, 50]. GLCM shows good results in situation where the textures are separable easily. In addition, various resampling techniques are applied to tackle the class imbalance. Since the inappropriate and ad hoc values of hyperparameters adversely influence the performance of a classifier, we have also optimized the hyperparameters of the random forest classifier by using a real encoded genetic algorithm. The results demonstrate that the performance of the RF classifier with the optimal configuration of hyperparameters and oversampling, as the class imbalance addressing technique, is significantly better than the decision tree and SVM classifiers. The results of the optimized RF classifier are also competitive to the deep learning approaches.

The paper is organized as given here: Section 2 describes applying machine learning techniques for Covid-19 diagnosis. Section 3 covers the proposed technique and the related methodology. Section 4 presents and discusses the results of the suggested approach. Finally, section 5 includes the conclusion and further enhancement of the research work.

2 Related Work

Quick and accurate Covid-19 diagnosis is essential to prevent its exponential spread and save as many humans lives as possible. The machine learning algorithms can play a vital role in assisting medical experts in early Covid-19 diagnosis. The section presents an overview of classification techniques for diagnosing Covid-19 cases predominantly from X-rays and CT scans.

Many machine learning algorithms have been envisaged to capture the patterns of spread of Covid-19 and identify the people infected by it [3, 5, 39, 63, 65]. An et al. [6] have developed predictive models for the prognosis of Covid-19 patients based on socio-demographic data, medical status, infection route, and history for the countrywide cohort of South Korea. They used many machine learning algorithms and their results showed that LASSO and linear SVM achieved reasonable sensitivities (90.7% and 92.0%, respectively) and specificities (91.4% and 91.8%, respectively). The study established that machine learning models can utilize socio-demographic and medical history data for predicting the prognosis of Covid-19 patients in the absence of any test reports. Alballa and Turaiki [4] have recently reviewed the Machine Learning (ML) approaches applied for Covid-19 diagnosis and mortality by using publicly available clinical and laboratory data. The authors have highlighted the use of imbalanced datasets prone to selection bias as one of the limitations of the existing ML applications for detecting Covid-19.

The performance of a disease diagnosis technique is not reliable in the face of class imbalance that frequently exists in the medical datasets. Confronting the class imbalance in the medical datasets is a vital question to prevent the bias of the classifier towards the majority class. The class imbalance is addressed by applying resampling techniques such as oversampling, undersampling, a combination of under and oversampling, Synthetic Oversampling Minority Technique

(SMOTE), Majority Weighted Minority Oversampling Technique (MWMOTE), and Random walk Oversampling (RWO) [10, 13, 70]. Jain and Ratnoo [32] have shown that the performance of oversampling technique has been satisfactory for dealing with the class imbalance present in disease diagnosis. A hierarchical and multi-class classification system is proposed to handle the imbalance class ratio of the datasets in [17].

In addition to RT-PCR and other clinical data, X-rays and CT scans medical images are valuable resources for diagnosing Covid-19 cases. Without any doubt, the RT-PCR test is the world over standard for detecting Covid-19 cases. However, since RT-PCR test machines are run with bunches of samples to diminish cost, the results are often unavailable on the same day. In extreme cases, medical experts may need to wait for two to three days before the RT-PCR reports are available. Further, RT-PCR testing is not a foolproof method and may have false negative rates ranging between 2% to 33% in repeat sample testing [64]. The consequences of false negative cases can be fatal for a disease like Covid-19 with high infection and mortality rates. A combination of RT-PCR and auto-diagnosing Covid-19 from medical images using ML can provide a reliable mechanism for containing the disease, starting early treatment, and preventing loss of lives.

Several researchers have implemented machine learning algorithms, comprising deep learning neural networks, for diagnosing Covid-19 from X-rays [12, 20, 46, 69]. Wang et al. [68] used RF with SMOTE as the resampling technique to predict the prognoses of Covid-19 positive cases. The authors could predict the death rate of Covid-19 cases with high accuracy. They also acknowledged two predicators, LDH higher than 500 U/L and Myo higher than 80 ng/ml, as a risk towards mortality rate. SVM-based deep feature extracting approach has been employed to identify Covid-19 patients through X-ray images [57]. The method achieved more than 95 percent accuracy for two chest X-ray datasets. Brunese et al. [12] envisaged a KNN technique to automatically discriminate between COVID-19 and other respiratory diseases by analyzing X-rays. Yoo et al. [69] used a decision tree classifier based on deep-learning for diagnosing Covid-19 from medical images. The authors trained three classifiers by a coevolutionary neural network based on the PyTorch framework. The first decision tree predicted the CRX images into normal and abnormal. The second decision tree captured the atypical images that contained indications of tuberculosis and the third one predicted the COVID-19 cases. The accuracies achieved by the three decision tree models were 98, 80 and 95 percent, respectively. Mostafiz et al. [46] proposed a hybridization of discrete wavelet transforms (DWT) feature extraction technique and deep coevolutionary NNs to capture Covid-19 cases from chest X-rays. The images were initially enhanced and segmented and, then features were extracted by applying deep CNN and DWT. A random forest-based bagging classifier was also used to detect the Covid-19 cases. The proposed approach attained an accuracy of 98.5 percent.

In addition to X-ray images, CT scan images have also been extensively used for detecting Covid-19 disease. Kadry et al. [35] have applied decision tree, random forest, and SVM classifiers to classify CT scan images for the presence and absence of Covid-19 induced pneumonia. The standard statistical and machine learning method aggregate, called a hybrid classifier system, has also been applied to extract features from the CT-scan images [22]. Machine learning algorithms such as logistic regression, neural networks and random forest have recently been applied to increase the performance of RT-PRC and chest-CT scans for diagnosing Covid-19. Machine learning models increased the AUC values of RT-PCR and Chest-CT from0.778 to 0.892 and from 0.852 to 0.930, respectively [23]. Ensemble methods, known for their prediction accuracy, have also been used for Covid-19 diagnosis [9, 62]. Random forest classifiers have been proposed for the large-scale screening and predicting thousands of covid-19 patients from CT scans [60, 66]. Zhou et al. [71] have investigated an ensemble-based deep learning approach for Covid-19 prediction from CT scans. Das et al. [16] have suggested a Deep CNN based approach to detect covid-19 cases using chest X-Rays. They combined three CNN models: DenseNet201, Inceptionv3, and Resnet50V2. Every model is trained independently, and finally, the result was combined using the average weighted ensemble technique.

Deep and transfer learning approaches, which involve deep neural networks with a large number of hidden layers, have been widely implemented for diagnosing Covid-19 patients from medical images [15, 33, 37, 41, 43, 45, 51]. The deep and transfer learning techniques have several advantages. These techniques automatically accomplish feature extraction, feature selection, and classification without human intervention. Further, these techniques usually achieve high accuracy. However, the advantages of deep learning are not without cost. The deep learning techniques need an enormous amount of training data and suffer from high computational costs for training phases. Moreover, deep neural networks are black-box classifiers and not very comprehensible for human reasoning and interpretation. Therefore, these techniques should only be applied for the complex and large classification problems which cannot be solved otherwise

In addition to traditional ML classification algorithms and deep learning approaches, evolutionary and other nature-inspired methods have also been applied for Covid-19 diagnosis using medical image data. Elaziz et al. [20] suggested a novel foraging optimization based on differential evolution in a parallel multi-core computational framework for classifying chest X-rays for detecting the presence and absence of Covid-19 disease. Shaban et al. [58] have used genetic algorithms as a wrapper approach for feature selection from CT scans. They have proposed an enhanced KNN classifier to detect covid-19 sufferers.

Hyperparameter settings influence the efficacy of classifier algorithms. Nevertheless, the research towards optimizing the hyperparameters of the classifiers to boost the Covid-19 detection rate is scarce. We could only find three such research works. First, He et al. [28] have investigated DenseNet3D121, an approach that extensively focuses on the process of hyperparameter tuning. The study achieved an accuracy of 88.63%, an F1-score of 88.14%, and an AUC of 94.0%. Second, Singh et al. [61] have investigated a convolutional neural network to classify the infected patients of the covid-19. In this study, the authors have tuned the preliminary parameters of

CNN through a multi-objective differential evolution scheme. A simpler and comprehensible machine learning technique is preferable (particularly for a domain like medical diagnosis where the reasoning behind the diagnosis is no less critical than the diagnosis itself) provided that it gives a satisfactory predictive performance. Hence, we use a random forest classifier model with its hyperparameters optimized through a genetic algorithm for Covid-19 detection. We also use resampling techniques to address the class imbalance before applying the classifier.

3 The Proposed Optimal RF Classifier System

This section describes the proposed optimal RF classifier for Covid-19 diagnosis. The section consists of the following major components: description of the datasets, feature extraction, resampling, partitioning the data, applying a genetic algorithm for optimizing the hyperparameters of the classifiers. The overall block diagram for the proposed system is shown in Fig. 1.

3.1 Datasets

This research uses X-ray and CT scan images for Covid-19 diagnosis. We have extracted the Covid-19 datasets from the Mendeley data repository [21]. The covid-19 dataset is a massive collection of X-ray and CT scan images. Total 9537 X-ray images were collected with the size of 512×512 pixels; out of these images, 4044 belong to the Covid-19 positive cases, and the remaining 5493 are non-Covid patients. Similarly, out of 8053 CT scan images of 512×512 pixels, 5426 images are of Covid-19 positive cases, and the remaining 2627 images are of non-Covid patients. Thus, the X-ray images are in a ratio of 42:58 for positive and negative cases, whereas the CT scan dataset has a ratio of 67:33 for the positive and negative covid-19 patients. Thus, there is a significant class imbalance in the CT scan dataset.

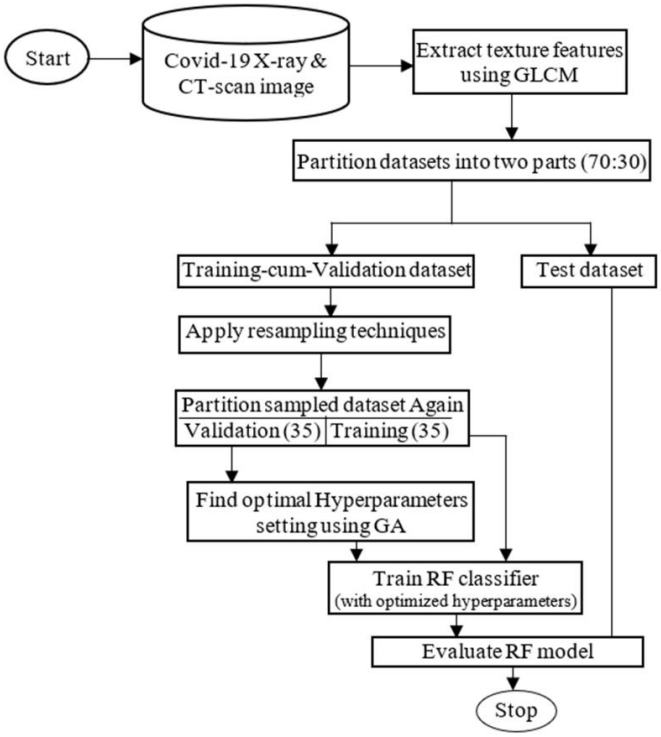

Fig. 1. The proposed Random Forest classifier system.

3.2 Feature extraction

The first step for image classification is features extraction. Haralick et al. [26] introduced the GLCM method for texture feature extraction. The GLCM method extracts features by finding Grey-Level Co-occurrence Matrix, called GLCM in short. In GLCM, the number of grey levels in the image corresponds to the number of rows and columns. The GLCM technique stores the specific dependencies of grey levels in an image [44, 67]. In the beginning, the value of each element in GLCM (i, j) is zero. Then the relationship between the pixels is computed horizontally towards the right. The value of each cell is updated according to the manifestation of the pixels together. We can compute texture features such as contrast, correlation, homogeneity, energy, etc., using GLCM [18].

We have used the GLCM method for feature extraction before implementing image classification tasks GLCM has extracted 22 features for each X-ray and CT scan image dataset. The attributes extracted from the GLCM matrix are assigned names automatically itself [29]. In addition, an appropriate class label is also given for every data tuple for the corresponding image (Covid-Positive and Covid-Negative). All the textural features extracted from the GLCM method are considered for final input to the classifiers.

3.3 Partitioning data and applying sampling techniques

The X-ray dataset is simply portioned in a ratio of 70:30 for training and test datasets. We have seen that the CT scan dataset has a significant class imbalance. Here, it is essential to

address the class imbalance problem before training a classifier for disease diagnosis [19]. For the CT-scan dataset, we have initially partitioned it into two parts-Training-cum-Validation of, and test datasets in a 70:30 ratio. Four resampling techniques- (Oversampling (Over), Undersampling (Under), a combination of Over and Undersampling (Over+Under), Synthetic Oversampling Minority Technique (SMOTE)) is applied on the Training-cum-Validation dataset. After applying to resample, the training-cum-Validation dataset is further divided into validation and training data in equal proportions. Subsequently, the RF classifier hyperparameters are optimized on the validation dataset. The step gives us optimal hyperparameter values. Now, the RF classifier is trained on the training data with its hyperparameters set to the optimal configuration, and finally, its performance is measured on the test data.

3.4 Building RF Classifier

This paper has employed the Random Forest (RF) algorithm developed by Leo Breiman [11]. RF is a tree-based classification algorithm known for its accurate predictions. RF classifier consists of a collection of several unpruned decision trees making a forest. The set of decision trees are generated from the random samples of the training data. Further, out of a total of M number of attributes, m attributes are randomly chosen for producing the best split at any node. For the whole process of random forest, the value of 'm' remains constant. Each tree is grown to the most significant possible extent. RF makes predictions by majority vote from the ensemble of trees. In addition to the RF classifier, we have used the Recursive Partitioning and Regression Tree, a variant of the decision tree, and support vector machine classifier deep learning approaches for comparison purposes.

3.5 Hyperparameters settings

We have selected the hyperparameters that directly impact the functioning of the random forest classifier. We have considered the number of trees, the number of attributes involved in splitting a node, and the maximum number of nodes in any tree in the random forest classifier. We have also optimized the hyperparameters of the Decision Tree and SVM classifiers for a fair comparison. Table 2 lists the hyperparameters of various classifiers and their range of values that are optimized.

3.6 Optimizing hyperparameters of RF

The hyperparameters are tuned for the random forest and other classifiers with the help of a genetic algorithm. GA is employed to return an optimal configuration for hyper-parameters for the classifiers. The GA uses a numeric encoding scheme, heuristic crossover, and random real mutation operators. The geometric mean of sensitivity and specificity of the RF classifier is taken as the fitness function for the GA. The parameters settings for the genetic algorithm are given in table 1.

Table 1. GA parameters setting.

Population Size	Number of Generation	Crossover Rate	Elitism	Mutation Rate
50	100	0.7	1	0.1

3.7 Tools used

We have implemented all the methods in R studio, which is an open-source tool for applying statistical and machine learning techniques. We have sourced the 'random forest' package to implement a random forest classifier, 'rpart' package to implement a decision tree. Furthermore, 'DMwR', 'Imbalance', and 'ROSE' packages are imported into the R script for addressing the class imbalance. Finally, the 'caret' package is used for the performance evaluation of the classifier. Hyperparameters configuration of the decision tree, random forest and SVM is given for the GA chromosome in Table 2.

Table 2. Hyperparameters configuration for the GA chromosome.

Random Forest		Decision tree		SVM	
Name of hyperparameter	**Range of Values**	**Name of hyperparameter**	**Range of Values**	**Name of hyperparameter**	**Range of Values**
# trees in the forest (ntree)	100:500	min split	10:100	Cost	0.1:4
# Attrib. for node split (mtry)	3: sqrt(#attrb)	complexity	0.01:0.3	Gamma	0.01:2
Maximum nodes (maxnode)	5:50	Maximum depth	1:20	Epsilon	0.01:2
---	----	Splitting criteria	0:1	-	-

3.8 Evaluating classifiers

Measuring accuracy is not a reasonable indicator of a classifier's performance for disease diagnosis. Therefore, it is essential to consider the sensitivity, specificity, geometric mean, and AUC as the performance metrics in addition to accuracy. Confusion metrics provides true positive (TP) true negative (TN), false positive (FP) and false negative (FN) cases for the calculation of the performance metrics. Confusion matrix is shown in Table 3. The caret package of R, statistical tool is used to evaluate the classifiers. The various performance metrics used in this study are defined below from "Eq. (1-5)".

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Geometricmean = \sqrt{Sensitivity + Specificity} \quad (4)$$

$$F - measure = \frac{TP}{TP+\left(\frac{FP+FN}{2}\right)} \quad (5)$$

Where TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative instances.

Table 3. Confusion matrix.

Actual	Predicted	
	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

3.9 Comparative study

We have experimented with combinations of various classifiers with sampling techniques and tunning of parameters through GA. The different experimental combinations and their short forms are listed below in Table 4.

Table 4. List of Machine Learning Algorithms.

Description of Machine Learning Algorithm	Short Form
Recursive Partition and Regression Trees	Rpart
Support Vector Machine	SVM
Random Forest	RF
Recursive Partition and Regression Trees + Genetic Algorithm for parameter tuning	Rpart+GA
Support Vector Machine + Genetic Algorithm for hyperparameter tuning	SVM+GA
Random Forest+ Genetic Algorithm for hyperparameter tuning	RF+GA
Recursive Partition and Regression Trees + Oversampling+Genetic Algorithm for parameter tuning	Rpart+Over+GA
Support Vector Machine +Oversampling+ Genetic Algorithm for hyperparameter tuning	SVM+Over+GA
Random Forest+ Oversampling+Genetic Algorithm for hyperparameter tuning	RF+Over+GA

4 Experimental Results

A random forest classifier with optimal hyperparameter configuration is applied to X-ray and CT scan images. The performance of the RF classifier is also compared with Rpart and SVM classifiers. The experimental results for various experiments conducted on X-ray images are shown in Table 5. All the results reported in this research are based on the average of the ten runs.

Table 5. Results for Covid-19 X-ray dataset.

ML Algorithms	Accu	Sens	Spec	Gmean	F1	AUC
Rpart	0.81	0.78	0.84	0.81	0.78	0.81
Rpart+GA	0.84	0.78	0.87	0.82	0.8	0.83
SVM	0.83	0.76	0.88	0.82	0.79	0.82
SVM+GA	0.85	0.8	0.96	0.88	0.81	0.86
RF	0.90	0.92	0.86	0.89	0.91	0.89
RF+GA	0.94	0.95	0.91	0.93	0.92	0.92

The results show that optimizing the hyperparameters of the classifiers through GA improves their performance. Further, the RF classifier with optimized hyperparameter gives the best performance. We have used four data resampling techniques for the CT scan dataset to resolve the class imbalance. The experimental results for the CT scan are shown in Table 6. However, results are listed only for the resampling techniques that lead to the best performance of the classifiers.

Table 6. Results for CT-Scan dataset.

Classification Method	Acc	Sens	Spec	Gmean	F1	AUC
Rpart	0.88	0.95	0.73	0.83	0.92	0.84
SVM	0.87	0.93	0.74	0.83	0.90	0.83
RF	0.93	0.87	0.96	0.91	0.90	0.92
Rpart+Over	0.88	0.87	0.89	0.88	0.88	0.88
SVM+Over	0.91	0.96	0.87	0.91	0.92	0.91
RF+(Over and Under sampling)	0.97	0.96	0.98	0.97	0.97	0.97
Rpart+GA	0.90	0.93	0.83	0.88	0.92	0.88
SVM+GA	0.95	0.97	0.90	0.93	0.96	0.94
RF+GA	0.91	0.94	0.85	0.89	0.93	0.89
Rpart+Over+GA	0.87	0.82	0.93	0.87	0.86	0.87
SVM+(Over and Under Sampling)+GA	0.96	0.98	0.95	0.96	0.97	0.96
RF+Over + GA	0.99	0.98	0.99	0.98	0.99	0.99

The results demonstrate that oversampling and a combination of oversampling and undersampling techniques worked best for addressing the class imbalance. Furthermore, applying the resampling technique enhances the performance of the classifiers, and optimizing the hyperparameters of RF classifiers boosting the performance of the Covid-19 diagnosis significantly.

4.1 Comparison with deep learning approaches

Deep learning techniques have been extensively suggested in the literature for Covid-19 disease diagnosis from medical images. Researchers have used different datasets and classification algorithms. The datasets, even from the same source, are not of the same size across the research papers. Therefore, an exact one-to-one comparison of the various techniques adopted for Covid-19 diagnosis in the various research works is difficult. We have selected a few nearest researchers for comparison. However, the results are only indicative. We

compare the proposed method for Covid-19 detection from X-ray images with the following two research works.

Rahimzadeh and Attar [54]: This study combined Xception and ResNet50V2 to extract multiple attributes from X-ray images. The authors collected X-ray images from two open-source repositories (from Kaggle and Github) containing 6234 chest X-ray images of Covid-19 positive persons. Their dataset contained three types of class labels: Normal, Pneumonia, and Covid-19. They also applied a novel training technique to balance the class skew.

Nour et al. [49]: This research work suggested convolution neural network architecture to auto-extract the discriminative features of chest X-ray images. The research used the existing characteristics of convolution neural networks like abstraction, filter family, and weight sharing to train the model from scratch. The deep feature extraction process was employed to provide input to machine learning methods (k- nearest neighbor, support vector machine, decision tree). The Bayesian optimization algorithm was used for tunning the hyperparameters. The comparison of the two above research works is summarized in Table 7.

Table 7. Comparison of the RF with deep learning approaches for X-ray images.

Model	Sensitivity	Specificity	Accuracy
Concatenated Model [54]	80.53	99.56	99.50
Deep Features + Bayesian Optimization [49]	89.39	99.75	98.97
RF+Over+GA	95.00	91.00	94.00

Table 7 shows that the RF classifier with optimized hyperparameters attains a significantly higher sensitivity than the two deep learning approaches. However, the specificity and accuracy of the proposed model are less. On the other hand, the high sensitivity means lesser false-negative cases, which is an essential criterion for disease diagnosis. The following three works are compared to the research presented in this paper for Covid-19 diagnosis from CT scan datasets.

(1) Goel et al. [24]: This study proposed a robust feature extraction framework using autoencoder (used for nonlinear layers to understand complex hierarchical to reconstruct input) and the GLCM function. The authors have employed the RF classifier to detect Covid-19 from CT scan images.

(2) Perumal et al. [52]: The researchers suggested a hybrid learning AlexNet+SVM based model for the COVID-19 classification of CT scan images. The authors worked on the datasets collected from Coronacase, Radiopaedia, Google images, and Github repositories.

(3) Polsinelli et al. [53]: This paper proposed a light CNN model based on the SqueezNet approach for Covid-19 recognition. It applied the Bayesian optimization method to optimize the hyperparameters for the CNN model with/without transfer learning. The datasets are collected from two different sources.

The comparison of the three existing research works with RF classifier with optimal hyperparameter setting is summarized in Table 7. RF classifier with optimal hyperparameters achieves significantly higher accuracy, sensitivity, and AUC than the three deep learning approaches, as shown in boldface in Table 8. However, the specificity and F-score of the proposed models are comparable [24].

Table 8. Comparison of the RF with deep learning approaches for CT scan images.

Model	Acc	Sens	Spec	Gmean	F1	AUC
Light CNN model [53]	0.85	0.88	0.82	0.85	0.86	-
Hybrid AlexNet+SVM [52]	0.97	0.96	98.0	0.88	0.97	0.93
Efficient Deep Network [24]	0.98	0.97	0.99	0.98	0.99	0.98
RF+GA+Over	0.99	0.98	0.99	0.98	0.99	0.99

Table 8 shows that the proposed system is competitive with deep learning approaches. Overall, experimental research presented in this paper indicates that an RF classifier in combination with resampling techniques and genetic algorithm to find optimal hyperparameters is a good option for Covid-19 diagnosis from X-Rays and CT scan images. Furthermore, such machine learning techniques can be used in combination with RT-PCR for enhancing the efficacy of Covid-19 diagnosis.

5 Conclusion

We have presented a Covid-19 diagnostic approach that uses a random forest classifier with hyperparameters optimized using a real-value encoded genetic algorithm. The proposed Covid-19 disease diagnosis method extracts the features from medical images through the GLCM technique. In addition, the problem of class imbalance is addressed by resampling techniques.

The results obtained from the current approach were found superior compared to the other related methods. Furthermore, the results were competitive to deep learning approaches, which take a lot of more computational resources. The main contribution of this research is to enhance the efficacy and efficiency of Covid-19 diagnosis using image datasets to supplement the results of RT-PCR tests. The proposed Covid-19 disease diagnosis method has successfully reduced the false-negative rate, essential for an infectious disease such as Covid-19. In the future, we intend to apply the deep forest for larger image datasets for further improvement in the Covid-19 diagnosis.

References

1. Abbas, A. et al.: Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. Appl Intell. (2020). https://doi.org/10.1007/s10489-020-01829-7.
2. Ahmad, A. et al.: Decision Tree Ensembles to Predict Coronavirus Disease 2019 Infection: A Comparative Study. Complexity, (2021). https://doi.org/10.1155/2021/5550344.
3. Albahri, A.S. et al.: Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel

Coronavirus (COVID-19): A Systematic Review. J Med Syst. 44, 7, (2020). https://doi.org/10.1007/s10916-020-01582-x.
4. Alballa, N., Al-Turaiki, I.: Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Informatics in Medicine Unlocked. 24, 100564 (2021). https://doi.org/10.1016/j.imu.2021.100564.
5. Alimadadi, A. et al.: Artificial intelligence and machine learning to fight COVID-19. Physiological Genomics. 52, 4, 200–202 (2020). https://doi.org/10.1152/physiolgenomics.00029.2020.
6. An, C. et al.: Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Scientific Reports. 10, 1, 18716 (2020). https://doi.org/10.1038/s41598-020-75767-2.
7. Apostolopoulos, I.D., Mpesiana, T.A.: Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 43, 2, 635–640 (2020). https://doi.org/10.1007/s13246-020-00865-4.
8. Ardakani, A.A. et al.: Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med. 121, 103795 (2020). https://doi.org/10.1016/j.compbiomed.2020.103795.
9. Bansal, V. et al.: AI-based Diagnosis of COVID-19 Patients Using X-ray Scans with Stochastic Ensemble of CNNs. (2020). https://doi.org/10.36227/techrxiv.12464402.v1.
10. Barua, S. et al.: MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. IEEE Trans. Knowl. Data Eng. 26, 2, 405–425 (2014). https://doi.org/10.1109/TKDE.2012.232.
11. Breiman, L.: Random Forests. Machine Learning. 45, 1, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.
12. Brunese, L. et al.: Machine learning for coronavirus covid-19 detection from chest x-rays. Procedia Computer Science. 176, 2212–2221 (2020). https://doi.org/10.1016/j.procs.2020.09.258.
13. Chawla, N.V. et al.: SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 1, 321–357 (2002).
14. Chung, M. et al.: CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology. 295, 1, 202–207 (2020). https://doi.org/10.1148/radiol.2020200230.
15. Civit-Masot, J. et al.: Deep Learning System for COVID-19 Diagnosis Aid Using X-ray Pulmonary Images. Applied Sciences. 10, 13, 4640 (2020). https://doi.org/10.3390/app10134640.
16. Das, A.K. et al.: Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network. Pattern Anal Applic. (2021). https://doi.org/10.1007/s10044-021-00970-4.
17. Di, D. et al.: Hypergraph learning for identification of COVID-19 with CT imaging. Medical Image Analysis. 68, 101910 (2021). https://doi.org/10.1016/j.media.2020.101910.
18. Ding, J. et al.: A Machine Learning Based Framework for Verification and Validation of Massive Scale Image Data. IEEE Transactions on Big Data. 1–1 (2017). https://doi.org/10.1109/TBDATA.2017.2680460.
19. Dittman, D.J. et al.: Comparison of data sampling approaches for imbalanced bioinformatics data. Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014. 268–271 (2014).
20. Elaziz, M.A. et al.: New machine learning method for image-based diagnosis of COVID-19. PLOS ONE. 15, 6, e0235187 (2020). https://doi.org/10.1371/journal.pone.0235187.
21. El-Shafai, W., Abd El-Samie, F.: Extensive COVID-19 X-Ray and CT Chest Images Dataset. 3, (2020). https://doi.org/10.17632/8h65ywd2jr.3.
22. Farid, A.A. et al.: A Novel Approach of CT Images Feature Analysis and Prediction to Screen for Corona Virus Disease (COVID-19). (2020). https://doi.org/10.20944/preprints202003.0284.v1.
23. Gangloff, C. et al.: Machine learning is the key to diagnose COVID-19: a proof-of-concept study. Sci Rep. 11, 1, 7166 (2021). https://doi.org/10.1038/s41598-021-86735-9.
24. Goel, C. et al.: Efficient Deep Network Architecture for COVID-19 Detection Using Computed Tomography Images. Radiology and Imaging (2020). https://doi.org/10.1101/2020.08.14.20170290.
25. Han, J. et al.: Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann (2011).
26. Haralick, R. et al.: Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on. SMC3, 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314.
27. Harmon, S.A. et al.: Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nature Communications. 11, 1, 4080 (2020). https://doi.org/10.1038/s41467-020-17971-2.
28. He, X. et al.: Benchmarking Deep Learning Models and Automated Model Design for COVID-19 Detection with Chest CT Scans. medRxiv. 2020.06.08.20125963 (2020). https://doi.org/10.1101/2020.06.08.20125963.
29. Humeau-Heurtier, A.: Texture Feature Extraction Methods: A Survey. 7, 26 (2019) https://doi.org/10.1109/ACCESS.2018.2890743.
30. Ismael, A.M., Şengür, A.: Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Systems with Applications. 164, 114054 (2021). https://doi.org/10.1016/j.eswa.2020.114054.
31. Iwendi, C. et al.: COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Front. Public Health. 8, (2020). https://doi.org/10.3389/fpubh.2020.00357.
32. Jain, A. et al.: Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach. In: 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC). pp. 1–8 (2017). https://doi.org/10.1109/ICOMICON.2017.8279150.
33. Jain, G. et al.: A deep learning approach to detect Covid-19 coronavirus with X-Ray images. Biocybernetics and Biomedical Engineering. 40, 4, 1391–1405 (2020). https://doi.org/10.1016/j.bbe.2020.08.008.
34. Jin, W. et al.: Hybrid ensemble model for differential diagnosis between COVID-19 and common viral pneumonia by chest X-ray radiograph. Computers in Biology and Medicine. 131, 104252 (2021). https://doi.org/10.1016/j.compbiomed.2021.104252.
35. Kadry, S. et al.: Development of a Machine-Learning System to Classify Lung CT Scan Images into Normal/COVID-19 Class. arXiv e-prints. 2004, arXiv:2004.13122 (2020).
36. Kanne, J.P. et al.: Essentials for Radiologists on COVID-19: An Update—Radiology Scientific Expert Panel. Radiology. 296, 2, E113–E114 (2020). https://doi.org/10.1148/radiol.2020200527.
37. Karthik, R. et al.: Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN. Applied Soft Computing. 106744 (2020). https://doi.org/10.1016/j.asoc.2020.106744.
38. Kong, W., Agarwal, P.P.: Chest Imaging Appearance of COVID-19 Infection. Radiology: Cardiothoracic Imaging. 2, 1, e200028 (2020). https://doi.org/10.1148/ryct.2020200028.
39. Lalmuanawma, S. et al.: Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals. 139, 110059 (2020). https://doi.org/10.1016/j.chaos.2020.110059.
40. Li, W.T. et al.: Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC

Medical Informatics and Decision Making. 20, 1, 247 (2020). https://doi.org/10.1186/s12911-020-01266-z.
41. Liang, S. et al.: Fast automated detection of COVID-19 from medical images using convolutional neural networks. In Review (2020). https://doi.org/10.21203/rs.3.rs32957/v1.
42. Mall, P.K. et al.: GLCM Based Feature Extraction and Medical X-RAY Image Classification using Machine Learning Techniques. In: 2019 IEEE Conference on Information and Communication Technology. pp. 1–6 IEEE, Allahabad, India (2019). https://doi.org/10.1109/CICT48419.2019.9066263.
43. Marques, G. et al.: Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. Applied Soft Computing. 96, 106691 (2020). https://doi.org/10.1016/j.asoc.2020.106691.
44. Mattonen, S.A. et al.: New techniques for assessing response after hypofractionated radiotherapy for lung cancer. J Thorac Dis. 6, 4, 375–386 (2014). https://doi.org/10.3978/j.issn.2072-1439.2013.11.09.
45. Minaee, S. et al.: Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Medical Image Analysis. 65, 101794 (2020). https://doi.org/10.1016/j.media.2020.101794.
46. Mostafiz, R. et al.: Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features. Journal of King Saud University - Computer and Information Sciences. S1319157820306182 (2020). https://doi.org/10.1016/j.jksuci.2020.12.010.
47. Muhammad, L.J. et al.: Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN COMPUT. SCI. 2, 1, 11 (2020). https://doi.org/10.1007/s42979-020-00394-7.
48. Narin, A. et al.: Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. arXiv:2003.10849 [cs, eess]. (2020).
49. Nour, M. et al.: A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization. Applied Soft Computing. 97, 106580 (2020). https://doi.org/10.1016/j.asoc.2020.106580.
50. Öztürk, Ş., Akdemir, B.: Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM and SFTA. Procedia Computer Science. 132, 40–46 (2018). https://doi.org/10.1016/j.procs.2018.05.057.
51. Pathak, Y. et al.: Deep Transfer Learning Based Classification Model for COVID-19 Disease. IRBM. (2020). https://doi.org/10.1016/j.irbm.2020.05.003.
52. Perumal, V. et al.: Prediction of COVID-19 with Computed Tomography Images using Hybrid Learning Techniques. Disease Markers. 2021, 1–15 (2021). https://doi.org/10.1155/2021/5522729.
53. Polsinelli, M. et al.: A light CNN for detecting COVID-19 from CT scans of the chest. Pattern Recognition Letters. 140, 95–100 (2020). https://doi.org/10.1016/j.patrec.2020.10.001.
54. Rahimzadeh, M.: A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. Informatics in Medicine Unlocked. 9 (2020).
55. Rodrigues, J.C.L. et al.: An update on COVID-19 for the radiologist - A British society of Thoracic Imaging statement. Clin Radiol. 75, 5, 323–325 (2020). https://doi.org/10.1016/j.crad.2020.03.003.
56. Rohila, V.S. et al.: Deep learning assisted COVID-19 detection using full CT-scans. Internet of Things. 14, 100377 (2021). https://doi.org/10.1016/j.iot.2021.100377.
57. Sethy, P.K. et al.: Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine. Int J Math, Eng, Manag Sci. 5, 4, 643–651 (2020). https://doi.org/10.33889/IJMEMS.2020.5.4.052.
58. Shaban, W.M. et al.: A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. 18 (2020).
59. Shakhovska, N. et al.: The Hierarchical Classifier for COVID-19 Resistance Evaluation. Data. 6, 1, 6 (2021). https://doi.org/10.3390/data6010006.
60. Shi, F. et al.: Large-Scale Screening of COVID-19 from Community Acquired Pneumonia using Infection Size-Aware Classification. (2020).
61. Singh, D. et al.: Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks. Eur J Clin Microbiol Infect Dis. 1–11 (2020). https://doi.org/10.1007/s10096-020-03901-z.
62. Singh, P.D. et al.: A Novel Ensemble-based Classifier for Detecting the COVID-19 Disease for Infected Patients. Inf Syst Front. (2021). https://doi.org/10.1007/s10796-021-10132-w.
63. Somasekar, J. et al.: Machine Learning and Image Analysis Applications in the Fight against COVID-19 Pandemic: Datasets, Research Directions, Challenges and Opportunities. Materials Today: Proceedings. (2020). https://doi.org/10.1016/j.matpr.2020.09.352.
64. Surkova, E. et al.: False-positive COVID-19 results: hidden problems and costs. The Lancet Respiratory Medicine. 8, 12, 1167–1168 (2020). https://doi.org/10.1016/S22132600(20)30453-7.
65. Swapnarekha, H. et al.: Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review. Chaos, Solitons & Fractals. 138, 109947 (2020). https://doi.org/10.1016/j.chaos.2020.109947.
66. Tang, Z. et al.: Severity Assessment of Coronavirus Disease 2019 (COVID-19) Using Quantitative Features from Chest CT Images. (2020).
67. Veeramuthu, A.: Brain Image Classification Using Learning Machine Approach and Brain Structure Analysis. Procedia Computer Science. 7 (2015).
68. Wang, J. et al.: A descriptive study of random forest algorithm for predicting COVID-19 patients outcome. PeerJ. 8, e9945 (2020). https://doi.org/10.7717/peerj.9945.
69. Yoo, S.H. et al.: Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging. Front Med (Lausanne). 7, (2020). https://doi.org/10.3389/fmed.2020.00427.
70. Zhang, H., Li, M.: RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion. 20, 99–116 (2014). https://doi.org/10.1016/j.inffus.2013.12.003.
71. Zhou, T. et al.: The ensemble deep learning model for novel COVID-19 on CT images. Applied Soft Computing. 98, 106885 (2021). https://doi.org/10.1016/j.asoc.2020.106885.

***