Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 3 (2024)

An Optimized Machine Learning Framework for Automatic Detection of COVID-19 Using CT-Scan and X-Ray Images

[1] Sivanagireddy Kalli, [2] Bhuvan Unhelkar, [2] Siva Shankar S, [4] Prasun Chakrabarti

[1] Post-Doctoral Researcher Information Systems and Decision Sciences, Muma College of Business, University of South Florida,8350 N. Tamiami Trail Sarasota, Florida, USA.

[2] Professor, Information Systems and Decision Sciences, Muma College of Business, University of South Florida, 8350 N. Tamiami Trail Sarasota, Florida, USA.

[3] Professor, Department of Computer Science and Engineering, KG Reddy College of Engineering and Technology, Moinabad, Telangana, India –501 504

[4] Professor Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India –313601

[1]sivanagireddykalli@gmail.com, [2] bunhelkar@usf.edu, [3] drsivashankars@gmail.com, [4] drprasun.cse@gmail.com

Abstract— Application of artificial intelligence (AI) techniques for accurate COVID-19 diagnosis on X-ray and CT-Scan imaging modalities is crucial due to the difficulties with the RT-PCR test. Researchers hold in high regard the development of computer-aided diagnosis Systems based on AI by utilizing X-ray images and CT scans for accurate diagnosis of COVID-19. In this work, an optimization-based machine learning frame work (OMLF) with different feature extraction technique is introduced to diagnosis of the COVID-19 in X-ray and CT scan images. The feature extraction techniques are to extract the features that help distinguish the COVID & non-COVID images. The extracted features are considered for denoting the images as vectors. These image vectors are trained with six different machine learning algorithms. The hyper parameters of are optimized using six algorithms in this work, among all the Random Forest (RF) with Fire Fly Optimization Algorithm fairly diagnoses COVID-19 from the CT &X-ray image directory.

Keywords— "COVID-19 Detection", "Machine Learning Algorithms", "RF", "CT scan & X-ray image", "Features Extraction".

I. INTRODUCTION

WHO report states that RT-PCR verification is required for all COVID-19 diagnosis [1]. The only material found in coronaviruses is RNA (ribonucleic acid), which must be changed into DNA (deoxyribonucleic acid) to be amplified and it is carried out by RT-PCR test for detection of virus. However, doing the RT-PCR test requires equipped laboratories and specialised equipment that are unavailable in the majority of nations and to produce test results takes 24 hours, which causes to further disease spreading from the infected person. An important part of minimising community spread and isolating positive cases ahead of time is early detection of the disease. It is possible that the test result wasn't accurate and need to redo the RT-PCR test or other relevant tests.

The lung region is the most prominent part affected by the virus, which is scanned using either X-ray or Computed Tomography (CT) produces images, those are typically taken into consideration when assessing the infection's severity [2]. One kind of medical imaging method utilized for diagnosing COVID-19 is X-ray imaging [3]. The benefits of X-ray imaging include minimal costs and little chance of radiation exposure that could be harmful to people's health. The detection of COVID-19 with this imaging method is a somewhat challenging task. However, due to a lack of improved understanding of the disease, even for expert radiologists, infection prediction from medical images has become a difficult assignment. To decrease the detection error in COVID-19 detection, CT-Scan imaging is employed. With its exceptional resolution and contrast, CT scan images are highly effective to diagnosis of COVID-19 in lungs [4]. Additionally, CT-Scan is utilised as a clinical feature for patients with COVID-19 disease. The CT scans of COVID-19 images showed significant pulmonary destructions like extensive consolidation and interstitial inflammation. Multiple slices are recorded during CT-Scan imaging to diagnose COVID-19 in patients. Due to the large amount of CT scan images, experts must diagnose COVID-19 with Exactness. Factors like ocular fatigue or an excessive number of patients to read CT scans could cause doctors to misdiagnose COVID-19.

Machine learning has the ability to independently uncover meaningful patterns and relationships from vast datasets. In the context of COVID-19 diagnosis, various ML algorithms have been extensively utilized, furthermore, automatic detection of COVID-19 disease is achieved with the use of deep and transfer learning techniques. Deep learning techniques attain high accuracy by extracting features automatically. However, these methods are unsuitable for human insight and analysis, and are computationally costly, and require a large number of training instances. Therefore, whenever the basic machine learning approaches are unable to attain the necessary degree of accuracy, deep learning techniques should be utilised.

In this work, we developed an optimization-based machine learning framework for COVID-19 prediction. In this work, an experiment was performed with six different ML algorithms. Because of the reasonable comprehensibility and accuracy, decision trees are a popular choice among the classical categorization systems. Many separate classifiers combined into one unit is called an ensemble classifier. In ensemble classifiers, the classifiers' judgments are combined in some way to categorise the newly added instances. One instance of an ensemble classifier is Random Forest (RF). A random forest (RF) classifier consists of numerous decision trees, each built from different subsets of training data and features. Using a resampling method called bagging, RF splits each node in the trees based on a randomly selected subset of available attributes.

Majority voting is one technique for finalizing the predictive decision of random forest. The suitable unknown function among the class attributes and dependent attributes is better approximated by the ensemble classifiers since they are constructed by applying local search techniques from several starting points. Because of this, ensemble classifiers like Random forests (RF) offer greater accuracy compared to individual classifiers. There is less number of applications for ensemble classifiers in diagnosis COVID-19 Images from CT & X-rays, which are frequently utilised to diagnose pneumonia in its early stages.

This paper is organized into seven sections. The sections are as follows: section 1 explains introduction to the work and Section.2 explains the existing approaches. The datasets and its details are presented in section 3. The evaluation measures of ML algorithms are explained in section 4. The proposed framework and the components used in the proposed frameworks are presented in Section.5. The section 6 explains the experimental values of proposed approaches on X-rays and CT scan image datasets. The section 7 concludes this work with possible extensions to this work.

II. REVIEW ON EXISTING APPROACHES FOR COVID-19 DETECTION

Medical imaging, particularly CT imaging, is the subject of numerous prior research among the various modalities utilised for diagnosis because of its availability and accuracy. Additionally, doctors benefit greatly from the automation of diagnostic procedures. Navid Ghassemi et al., [5] uses a pre-trained DNN based strategy that achieves best accuracy of for the task at hand by utilising a Cyclic Generative Adversarial Net (Cycle GAN) model for data augmentation. The experiment was conducted with a dataset of 189 patient images of 3163 that has been gathered and labeled by doctors, in contrast to previous datasets, normal data from individuals suspected of having the disease of COVID-19 data, and it is open to the public.

The deep learning models with medical image analysis obtained more accurate and faster results for COVID-19 identification. H. Mary Shyni et al. reviewed [6] the most recent deep learning methods extensively like CNN algorithm for COVID-19 diagnosis. In this context, a key strength of deep learning is its ability to identify subtle and complex visual features within medical images. This can be difficult to detect even for experienced human experts. This will ultimately lead to more accurate and reliable diagnosis [7].

The author Sohaib Asif, Ming Zhao et al., [8] was introduced a rapid generate lightweight CNN architecture with excellent efficiency for identifying individuals infected with COVID-19. Initially, they used pre trained DNN. Secondly, a lightweight shallow CNN architecture is developed for classify the images with a low rate of false negatives.

When compared with other deep learning models, the proposed model utilizes lesser parameters and lesser complex.

Sara Hosseinzadeh Kassania et al., compared [9] well-known frameworks for automatic COVID-19 classification that rely on deep learning based techniques for feature extraction. Among a pool of deep CNNs such as VGGNet, NASNet, InceptionResNetV2, ResNet, InceptionV3, Xception, MobileNet, and DenseNet were selected to yield the most relevant features which is a crucial aspect of learning. The Bagging tree classifier with feature extractor method of DenseNet121 attained best classification accuracy of 99%. The combination of LightGBM with feature extractor method of ResNet50 attained second best classification accuracy of 98%.

Karrar Hameed et.al. [10] Developed a DNN autoencoder with CNN and Three CT image approaches are applied in this system after minor modifications to the categorization normal and COVID-19 cases. Its final performance was reported. According to experimental results, the CNN model produced the best accuracy rate, additionally, utilising CT scans, and the proposed approach has shown better performance than the popular models currently in use.

The purpose of Mundher Mohammed Taresh et al., [11] work is to assess how well the most advanced pre-trained CNNs perform automatic COVID-19 diagnosis using chest X-rays (CXRs). The Investigation is built on various pre-trained deep learning models; these algorithms are fine-tuned to optimize accuracy.The MobileNet and VGG16 achieved the best accuracy of 98.28%. However, the VGG16 shows best performance than all other approaches.
Efficient diagnosis and treatment planning mainly depends on the rapid and accurate segmentation of COVID-19 infection regions from CT scan images. S. K. Towfek et al., proposed [12] Covid-19 infection segmentation by using machine learning algorithms. The proposed method uses CNN architecture to extract meaningful features from images. Then, these features are trained with a segmentation model and this model is used for delineating the infection regions accurately by utilizing the combination of attention and U-Net mechanisms. The experiment was conducted on a variety of datasets including COVID-19 patient's CT scan images. They observed that the proposed method is superior in segmenting infected regions accurately by obtaining an average Jaccard index of 0.88 and a Dice coefficient of 0.92.
However, the poorer performances offered by handcrafted image processing algorithms are the biggest drawback. Nabila Mansouri et al., proposed [13] a solution first time for extracting COVID-19 features effectively from the images of Computed Tomography (CT) by using deep learning techniques. Since the COVID-CT-Dataset contains a small number of patients, the CNN model cannot be further trained to improve performance. As a result, the proposed method functions as a pipeline structure with two stages such as A CNN model is used for baseline classification and distances to feature vectors of CT image portions are used to re-rank the baseline results.

Zhang F developed [14] an application, where they gave direction by summarising the most recent advancements in the artificial intelligence based application of COVID-19 lung imaging and offering scientific methodologies for image recognition, evaluation, and segmentation.

Because chests X-rays (CXR) are easy to use and reasonably priced, they has become a popular screening option for lung-related disorders. Radiologists find that because CXR is inexpensive and provides results quickly, interpreting images is simple and affordable. P. V. Naresh et al., presented [15] a novel method to enhance the identification and categorization of Covid-19 by the examination of the chest X-rays (CXR). The proposed model builds a Weighted Average Ensemble Model by using pre-trained architectures such as InceptionV3, ResNet50, and VGG16. Even with the dataset limitations, the proposed model achieves a 98.33% accuracy rate. The outcomes validate the capability of weighted average ensemble models to aid in the identification of COVID-19 via categorization based on chest X-rays.

Using CT and X-ray images to use machine learning techniques has made it easier to diagnose COVID-19 accurately. Hossein Mohammad-Rahimi et al., analysed [16] analysed the performance of studies that employed machine learning and DL techniques on chest X-ray and CT scan images for COVID-19 diagnosis. These methods' accuracy ranged from 76% to over 99%, which denotes that DL and ML techniques is used to the COVID-19 clinical diagnosis. Ahmad Mozaffer Karim et al., proposed [17] DNN tool for COVID-19 identification. This article presents a new technique that combines Multiclass NB, CNN with ALO (Ant Lion Optimisation) Algorithm to process symmetric X-ray data. Additionally, CNN is integrated with a number of other classifiers, including DT, KNN, SVM and Softmax. They assessed and showed these classifiers' promising results were obtained by the CNN and NB classifier with ALO Algorithm, which had the shortest execution times and 98.25% F1-score, 100% precision, and 98.31% accuracy.

The automated method can help health professionals make timely diagnoses. This may help reduce the burden on the health care system and improve patient outcomes. M.A.Khan et al. study explores the potential of using CT scans and X-ray images for automated COVID-19 detection, reviewing existing research. and discuss challenges and future directions in this area [30].

Yassine Meraihi et al., presented overview [19] for predicting COVID-19 cases. They collected research papers from many articles. The articles are categorized into two classes as Deep Learning based approaches and Supervised Learning based approaches based on the analysis of articles. A description of the applied machine learning algorithm and a list of used parameters are provided for each category. The set of parameters for every used algorithm is specified by using various tables. They presented different details of research articles such as variety of problems addressed (diagnosis, detection, or prediction), the kind of datasets analysed (clinical data, time series, CT images, X-ray images, or text), and performance evaluation metrics used. The study presents a picture of the popular works through a number of statistics and a discussion of the data gathered.

Vikas Kumar et al., [20] proposed a new method to diagnose COVID-19 through a Grad-CAM visual explanation for the expected images. They attained the best AUC score of 98% by using the feature vector generated by the architecture of ResNet50 and TWSVM as classifier. Based on AUC, they observed that the feature vector generated by the ResNet50 performs better than any other CNN architecture.

III. DATASET CHARACTERISTICS

The COVID-19 datasets were gathered from Mendeley data repository [21]. Which are used in this work to diagnose the disease of COVID-19. The dataset includes a large collection of CT scan and X-ray images categorized into two classes: COVID and Non-COVID. It features augmented data created using various augmentation techniques, resulting in a total of 17,099 images across CT scans and X-rays. Specifically, the X-ray dataset comprises 9,537 images, each with a resolution of 512x512 pixels, organized into two separate folders. Among these, 4,044 X-ray images correspond to COVID-19 positive cases and are stored in one folder, while the remaining 5,500 images, representing Non-COVID cases, are stored in the other folder. Figure 1 Shows representative Test Images and Table 1 provides a detailed overview of the characteristics of the X-ray dataset.

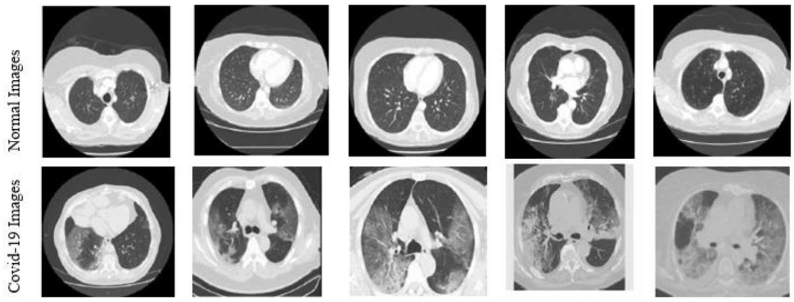

Fig. 1 The proposed study evaluated a collection of representative test images.

TABLE 1: DATASET CHARACTERISTICS OF X-RAYS IMAGES DATASET

Classes / Features	**COVID**	**Non-COVID**
Number of X-rays	4044	5500
Ratio Percentage of X-rays	42	58

Similarly, the CT scan images dataset contains a total of 8053 CT scan images with a pixel size of 512x512. Out of these 8053 CT scan images, 5426 CT scan images are associated with positive and remaining are negative of COVID-19. The dataset characteristics of CT scan images are presented in Table 2. As a result, the CT scan dataset has a notable class imbalance.

TABLE 2 DATASET CHARACTERISTICS OF CT SCAN IMAGES DATASET

Classes / Features	**COVID**	**Non-COVID**
Number of CT scan images	5427	2628
Ratio Percentage of CT scan images	67	33

Deep learning models have achieved great results in areas like computer vision natural language processing, and speech recognition. This success comes from their ability to learn complex patterns from big datasets. To train these models and , it's crucial to split the dataset into three parts: the training set, validation set, and test set. The model learns to do its job using the training set. Then, the validation set helps to pick the best model and adjust its settings during training . After the model has finished training with both the training and validation data, we use the test set to see how well it performs. This gives us an unbiased look at how the model will work on new unseen data .

IV. EVALUATION MEASURES

The efficacy of proposed methods in image classification is assessed by the application of several categorization algorithms. These methods use performance metrics like, recall, precision, and F1 Score, to display the results. Various metrics took into account of data assess the effectiveness. Confusion Matrix (CM) of the data True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) needed to determine the assessment measures. The Confusion Matrix displays class labels in the binary form, meaning that only two class labels—positive and negative—

are displayed on each axis. The dataset's classes has all been given positive and negative labels, with one class label being regarded as positive and the others as negative.

The performances measures are as fallows

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

V. PROPOSED OPTIMIZATION BASED MACHINE LEARNING FRAMEWORK (OMLF)

The proposed Optimization based Machine Learning Framework (OMLF) is represented in Fig. 2.

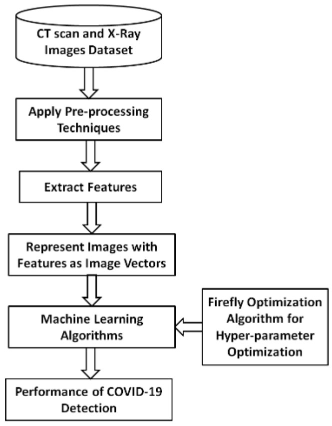

Fig.2. Proposed Optimization Approach

The suggested method starts by collecting CT scan and X-ray image datasets. Next, we use suitable pre-processing methods to get the images ready for the following phases. The key next step is to pull out important features from the images that can tell COVID cases apart from non-COVID ones. After we get these features, we use them to make feature vectors that stand for the images. We then use these vectors to teach six machine learning algorithms, which help us build the classification model. To check how well the machine learning algorithms work, we look at the hyperparameter values we used during the test.

The firefly optimization algorithm is used for determining the hyper-parameters and this algorithm is used for optimizing the hyper-parameter values. The ML algorithm's performance for COVID-19 detection is presented in Accuracy F1-Score, Precision, and Recall.

A. Feature Extraction Techniques

Feature extraction is the primary step in image classification. The Grey-Level Co-occurrence Matrix GLCM technique is used for extracting texture features was first presented by Haralick et al., [22]. According to GLCM, the number of columns and rows are matched to the count of grey levels present in the image. A GLCM method stores the particular grey level dependencies present in an image. Every element of GLCM (i, j) has a zero value at first. Subsequently, the horizontal relationship between the pixels is calculated in a rightward direction. Each cell's value is modified in accordance with how the pixels are manifested together. GLCM allows us to determine the texture features including energy, homogeneity, correlation, and contrast [23]. Prior to beginning of image classification tasks, we extracted features by using the GLCM approach. For every dataset of CT scan and X-ray images, GLCM has retrieved 22 features. Names are automatically allocated to the attributes that are extracted from the GLCM matrix [24].

Furthermore, each data tuple for each matching image (COVID and Non-COVID) is assigned the proper class label. For the final input to the classifiers, every textural feature that was retrieved by using the GLCM approach is taken into account. In order to apply traditional machine learning techniques for image classification issues, feature extraction is a necessary step. A media filter was applied to images so as to highlight a number of extra features that were included as extra numerical attributes. From the "dataset of CT scans and X-ray images", the Histogram of Oriented Gradients (HOG) feature was retrieved for this work. The HOG is a tool used by most image processing systems to characterise local geometric shapes inside an image based on the edge directions distribution. Because of HOG's robustness against noise and outliers, and invariant to tiny deformations, we chose HOG for alternative local shape descriptors. HOG is an image feature descriptor which is useful for ML and computer vision applications. In this work, we consider the Pyramid HOG (PHOG), which is a newly updated HOG that considers the local shape's spatial property when describing an image. In this experiment, resize the images to 64 by 128 pixels, which is the most frequently utilised dimension of image for the HOG. The images were then separated into cells at various levels of pyramid. The Equations (5) and (6) are used to determine the orientations and gradient magnitudes, where G_x and G_y represent the gradients of the horizontal and vertical respectively:

$$\theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{5}$$

$$GradientMagnitude = \sqrt{G_x^{\ 2} - G_y^{\ 2}} \tag{6}$$

The next step is to quantize each gradient orientation into K bins. Each pyramid resolution generates its HOG (Histogram of Oriented Gradients) vectors that are concatenated to form the final PHOG descriptor of the image. This concatenation contains the image spatial information. At all levels, the gradients of all the pixels in each cell are accumulated to create a local histogram of K bins. This leads to the final PHOG vector having a d dimension being formed by combining all the cells across different levels,

Equation (7) is used to determine the dimension d.

$$d = \sum_{l=0}^{L} 4^l \tag{7}$$

In this experiment, quantize the HOG descriptor into angle cases of 30 with 0 to 360 ranges. For the purpose of training our machine learning algorithm, K=30 and L=3 were chosen, & PHOG descriptor was used to extract 630 features. Estimating the features allows for additional analysis once the image is segmented. These features include the interpretation of images and create feature vectors that are used to categorise input datasets. These improve the rate of recognition. In this work, we extracted different features such as 10 features of the discrete wavelet transform, 32 features of textual and 36 features of statistical features from the image datasets to enhance the accuracy of COVID-92 classification. Discrete wavelet transforms include noise removal from the frequency domain and data compression. This comprises various components such as HH (high-high), HL (high-low), LH (low-high), and LL (low-low) that are decomposed by using wavelet. The LL component has the most information and energy. The textual features include correlation, entropy, homogeneity, contrast, and energy, which denote a second-order probability measures. The gray-scale distribution illustrates the image texture's periodicity for certain parameters such as sequential which is a sequence that is not generated randomly or uniformly. The first-order probability measures are represented by the statistical features, which include skewness, inverse difference moment, kurtosis, RMS, mean, standard deviation, smoothness, and variance. This work makes use of the defective portion's pixel histogram intensity. In this work, the final feature vector for images contains different features such as the GLCM technique of extracted features, HOG feature, DWT, statistical analysis, and textural analysis features, which are combined to incorporate the image's class (COVID or Non-COVID).

B. Machine Learning Algorithms

In a classification system, there are two major processes such as training of an ML algorithm and testing of a model. In the training process, the training dataset is considered for giving training to the ML algorithm. The trained algorithm is used for evaluating another set of instances which are called testing sets. Every instance is designated as a vector by using a set of feature values which are categorical or numeric in the training process. The classification performance and classifier learning time primarily depend on the feature set used in representation of the vector. Cross Validation (CV) concept is used for dividing the dataset into testing datasets and training datasets.

K-fold cross-validation tweaks standard cross-validation methods. It splits the whole dataset into k parts of the same size at random. This split makes sure each part has about the same mix of classes as the full dataset. In this approach, each part serves as the test set one time, while the other parts train the model. So, the classification algorithm runs k times giving k different accuracy scores from k separate tests. An accuracy of the classification algorithm is the average of k number of accuracies. The over-fitting is one general problem occurred in the training process of the classification algorithm. Obtaining high accuracy for testing is the primary aim of a classification algorithm.

In this work, six ML algorithms such as RF, SVM, LR, DT, NB, and KNN are utilized for the training of the datasets, which generates the model for finding the proposed approach performance.

1) K-Nearest Neighbour (KNN) Classifier

This is called a lazy learning algorithm because this classifier does not doing anything when there is no new instance for testing [25]. KNN algorithm remembers all instances in the training dataset as a substitute of generating any classification model. In KNN, the new instance class label is determined by comparing it with all instances in the training dataset. In this process, first, it computes the distance among all instances of the dataset and the new instance by using a distance measure. Generally, the KNN classifier uses Euclidian distance for continuous data and the Manhattan distance measure for discrete data for computing the distance. Next, KNN determines the *k* closest samples in the training dataset to a new instance by calculating the distance between them.

The K value in KNN classifier is a small integer which is defined by the user. The class label of new sample is decided based on the class label of highest number of instances in the K nearest neighbours. Because of the simplicity of KNN, most of the real word problems used this algorithm for classification.

2) Decision Tree (DT)

One non-parametric learning technique used in data mining is the Decision Tree classifier, which is applied to numerical, categorical, or a combination of both types of data [26]. To map instances to appropriate class labels, DT builds a classification model based on tree structures. Every internal node in a DT denotes a feature that is utilised to make decisions regarding the specific instance. The internal nodes and the leaf, which are designated as the result of a feature, are connected by the arcs. In the tree, the class label is indicated by each leaf. The path taken or satisfied from the root to a leaf in the tree is used to find the class label of a previously unknown sample; the appropriate class label of the leaf is then assigned to the new instance.

3) Naive Bayes (NB) Classifiers

The Bayesian classifier learned a probabilistic model to detect the unknown instance’s class label. In Bayesian classifiers, there is an assumption that the features are conditionally not related to the specific class label and the interactions among class labels and features were explained in terms of probability distributions. The Bayesian classifiers induce the class label’s probability distribution as well as every feature’s conditional probability distributions that are specific to a given class label. Then, these two Probability distributions are used to determine the class label of an unknown instance. [27]. Naïve Bayesian classifier is straightforward and common Bayesian classifier and is efficient in several problems that are faced in real-world and it violates the assumption of Bayesian classifiers because of complex relationships among features.

Logistic Regression

The categorization technique known as "logistic regression" has its roots in the statistical community [28]. Because LR requires less computing power, it is utilised to solve a variety of real-world issues. It is a general technique for resolving binary classification issues. Similar to linear regression, the primary goal of LR is to ascertain the weight of each input coefficient. The logistic function, known as the transformation function, is the foundation upon which LR operates. The logistic function predicts the class label by applying rules or probability. Using this function, numbers between 0 and 1 are converted. The performance of the LR classifier is enhanced by eliminating correlated or non-interacting features from the output.

4) Support Vector Machines (SVMs)

SVM was founded in 1963 by the authors Vapnik et al., for solving binary classification problems. Later, they extended the functionality of SVMs for adapting to problems of multi-class classification [29]. Support vector machine was chosen as the stable method because SVM is a flexible learning algorithm and hence has found its use in many real learning problems. The primary aim of SVMs is constructing or building one or more hyperplanes for dividing the specific dataset into more subsets consistent to various labels of classes. Here is a possibility of several hyperplanes that divide the dataset into subsets the SVM selects the hyperplanes which maximize the distances among the selected hyperplane and nearest instances in the training dataset. The hyperplane equations are formulated as an optimization problem, for instance, $w^T x+b \leq -1$ for class B and $w^T x+b \geq 1$ for class A.

Random Forest (RF)

Breiman created the Random Forest classification method [30] to average the choices made by different decision trees in order to lower variation. Random Forest (RF) (Breiman 2001) is an ensemble based machine learning method used for both regression and classification tasks. In order to mitigate the risk of overfitting, it creates several separate decision trees (DTs) with various subsets of the data points and characteristics. Once all the decision trees have made their prediction, in RF, the average value for regression tasks or the majority class (mode) for classification tasks is taken.

In this work, the authors concentrated on the classification portion of the algorithm for image classification. As previously stated, RF circumvents the widespread issue of over-fitting, which is accelerated by DTs, by applying standard technique of bootstrap aggregating or bagging to the Decision Tree learners.

C. Fire Fly Optimization Algorithm

In this work, we considered one of the popular optimization algorithms known as the FireFly Optimization Algorithm (FFOA) for optimizing the hyper-parameters in the training process of the deep learning algorithm [31]. The FFOA method is mainly inspired by the social behaviour and flashing characteristics of fireflies. The FFOA metaheuristic is the computation due to the intricate and complicated nature of the real-time firefly ecosystem. When implementing FFA, one crucial issue that must be resolved is the genuine formula of attraction. Generally speaking, firefly attraction is based on brightness, which in turn depends on the decision criterion. Equation (8) is used to evaluate the firefly's brightness at the x point in the minimization issue.

$$I(x) = \begin{cases} \dfrac{1}{f(x)} & if\ f(x) > 0 \\ 1 + f(x) & Otherwise \end{cases} \tag{8}$$

In Equation (8), I(x) denotes the attraction, while f (x) describes the decision criterion's value at the position of x. It is clear that the fitness function is represented by Equation (9).

$$I(r) = \frac{I_0}{1+\gamma r^2} \tag{9}$$

In Equation (9), I(r) denote the light intensity, r indicates the distance, and I_0 is the light intensity at the position of the source. Moreover, light appears weaker because it is somewhat absorbed by the surrounding air. These kinds of occurrences are illustrated by the coefficient of light absorption γ. According to the study, almost all FFOA implementations took into account how the inverse square law affected light absorption and distance by utilising the Gaussian formulation which is represented in Equation (10).

$$I(r) = I_0 \cdot e^{-\gamma r^2} \tag{10}$$

The relative force of attraction (β) of each firefly is affected by the distance between the target and the observer's firefly. Moreover, the force of attraction is directly proportional to the light intensity. This is calculated using equation (11)

$$\beta(r) = \beta_0 \cdot e^{-\gamma r^2} \tag{11}$$

In Equation (11), the attraction at distance r = 0 is denoted by the β_0 variable. Furthermore, it helps to identify the distance $\gamma = 1/\sqrt{\gamma}$ at which firefly attraction changes dramatically from β_0 to $\beta_0 e^{-1}$. It is commonly replaced with the Equation (12).

$$\beta(r) = \frac{\beta_0}{1+\gamma r^2} \tag{12}$$

Equation (13) denotes the movement of a subjective firefly i towards the new position (in the direction of t + 1 subsequent iteration) of the attractive firefly *j* brightness.

$$x_i^{t+1} = x_i^t + \beta_0 \cdot e^{-\gamma r_{ij}^2}\left(x_j^t - x_i^t\right) + \alpha^t\left(k - 0.5\right) \tag{13}$$

In Equation (13), α represents a random variable, rij is the distance among i and j, the parameter β_0 denotes the attraction in the direction of r = 0, and k is an arbitrary number generated from a uniform or Gaussian distribution. By comparing and updating every pair of firefly in all iterations, the firefly position is updated successively. The Cartesian distance is used to evaluate the distance between i and j, which is represented in Equation (14).

$$r_{ij} = \left\|x_i - x_j\right\| = \sqrt{\sum_{k=1}^{D}\left(x_{i,k} - x_{j,k}\right)^2} \tag{14}$$

In Equation (14), D denotes the number of parameters. For this problem, $\beta 0 = 1$ and $\alpha \in [0, 1]$ are appropriate values. A Fitness Function (FF) is created by the FFOA technique in order to improve classifier performances. In order to show which solution candidate performs best, it solves a positive integer. In these situations, Equation (15) treats the minimalized classifier error rate as FF. The optimal method achieves a lower error rate, whereas the worst-case scenario sees an increase in error rate.

$$FF\left(x_i\right) = ClassifierErrorRate\left(x_i\right) = \frac{number\ of\ misclassified\ samples}{Total\ number\ of\ samples} * 100 \tag{15}$$

VI. EXPERIMENTAL RESULTS

This work conducted experiments using various machine learning algorithms. To predict symptoms of COVID-19 from x-rays, CT scans, and image data sets. In this work six different machine learning algorithms, including KNN, DT, NB, LR, SVM, and RF with FFOA Classifier were used to evaluate the performance of the proposed approach for COVID-19 detection. Using the x-ray data set, the total results of the learning algorithm, are shown in Figure 3 & the performance values are presented in Table 4,

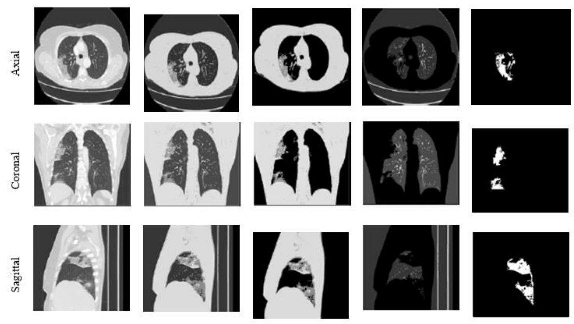

Fig. 3 OMLF results for CT scan images. (a) Test image, (b) Threshold image, (c) Threshold > Th, (d) Threshold < Th, (e) COVID-19 infection

TABLE 4 THE PERFORMANCE OF DIFFERENT ML ALGORITHMS FOR COVID-19 DETECTION

Classifier / Evaluation Measures	Recall	Precision	F1-Score	Accuracy
K-Nearest Neighbour	86.09	87.07	86.57	88.56
Decision Tree	89.90	90.48	90.19	91.66
Naïve Bayes	92.42	94.06	93.23	94.22
Logistic Regression	95.57	96.14	95.86	96.48
Support Vector Machine	98.76	98.94	98.85	99.02
Random Forest	99.41	99.50	99.46	99.54

Table 4 shows the performance indicators of the different classifiers for the detection of COVID-19. The Random Forest (RF) with FFOA classifier scored 99.41 for recall, 99.50 for precision, 99.46 for F1 score, and 99.54 for precision. The Support Vector Machine (SVM) classifier scored 98.76 for recall, 98.94 for Precision 98.85 for F1 score and 99.02 for, accuracy. The logistic regression (LR) classifier recorded scores of 95.57 for recall, 96.14 for precision, 95.86 for F1 score, and 96.48 for precision. The Naive Bayes (NB) classifier achieved scores for recall, precision. With F1 and accuracy scores of 92.42, 94.06, 93.23, and 94.22, respectively, the Decision Tree (DT) classifier reported a score of 89.90. for recall, 90.48 for precision, 90.19 for F1 score, and 91.66 for precision. Finally, the K-Nearest Neighbor (KNN) classifier scored 86.09 for recall, 87.07 for precision, 86.57 for F1 score, and 88.56 for Accuracy Figure 2 shows a comparative analysis of the performance of these machine learning algorithms in detecting COVID-19.

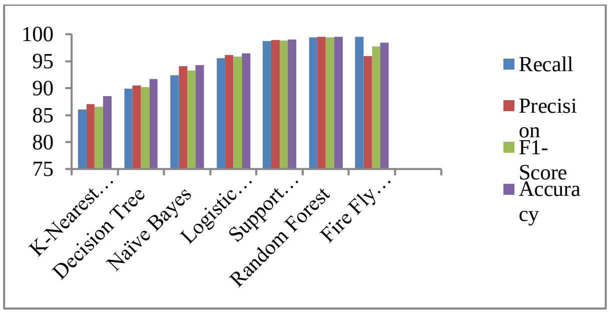

Fig. 4: Comparative Analysis of different algorithms for COVID-19 detection from X-Rays dataset

Figure 4 shows that the Random Forest algorithm with FFOA classifier achieved the highest performance scores in recall, accuracy, F1 score, and accuracy of COVID detection from the X-ray image dataset, outperforming than the other Machine Learning Algorithms In contrast, the K-Nearest Neighbor (KNN) classifier evaluated in this work shows the lowest performance score.

Table 5 shows the Consolidated Results of All ML Algorithms for COVID-19 Detection on CT scan images Dataset

TABLE 5 THE PERFORMANCE OF DIFFERENT ML ALGORITHMS FOR COVID-19 DETECTION ON CT SCAN DATASET

Classifier / Evaluation Measures	**Recall**	**Precision**	**F1-Score**	**Accuracy**
K-Nearest Neighbour	87.59	92.57	90.01	86.16
Decision Tree	91.24	92.68	91.96	89.07
Naïve Bayes	92.81	94.66	93.72	91.46
Logistic Regression	95.79	96.50	96.14	94.78
Support Vector Machine	96.73	97.71	97.22	96.23
Random Forest	97.59	98.52	98.06	97.37

From the Table. 5, measuring the performance of different types of classifiers for detecting COVID-19 disease It can be summarized as follows: Random forest (RF) algorithm with FFOA classification achieved recall, precision, F1 score, and precision scores of 97.59, 98.52, 98.06, and 97.37, respectively. Supported Vector Machine (SVM). The classifier recalled 96.73 in accuracy, 97.71, 97.22 in F1 score, and 96.23 in precision were recorded. The logistic regression (LR) classifier scored 95.79, 96.50, 96.14, and 94.78 for recall, precision, F1 score, and precision. The Nev Bayes (NB) classifier scored 92.81 for recall, 94.66. for accuracy, 93.22 for F1 score and 91.46 for accuracy. The Decision Tree (DT) classifier achieved recall, precision, F1 score, and precision values of 91.24, 92.68, 91.96, and 89.07, respectively. Finally, the K-Nearest Neighbor (KNN) classifier reported recall, precision. F1 and accuracy scores were 87.59, 92.57, 90.01, and 86.16, respectively. Figure 3 provides a comparative analysis of the performance of these machine learning algorithms in COVID -19. Detection

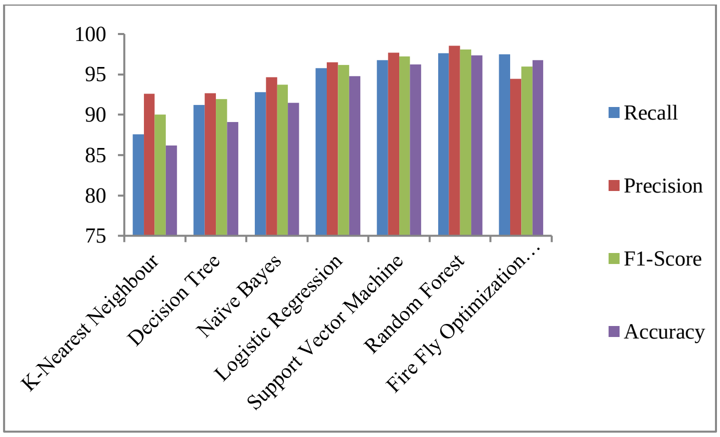

Fig.5. Comparative Analysis of different algorithms for COVID-19 detection from CT scan images dataset

Figure 5 shows that the Random Forest algorithm with FFOA classifier achieved the highest performance scores in recall, accuracy, F1 score, and accuracy of COVID detection from the CT scan image dataset, outperforming than the other Machine Learning Algorithms In contrast, the K-Nearest Neighbor (KNN) classifier evaluated in this work shows the lowest performance score.

VII. CONCLUSIONS &FUTURE SCOPE

This work introduces a new approach called an adaptive machine learning framework for COVID detection using X-ray, CT scan, and image datasets. The proposed method uses two image datasets to identify In the case of chronic obstructive pulmonary disease Feature extraction techniques were used to obtain relevant features that could effectively differentiate between COVID and non-COVID images. These extracted features are then used to be visualized as feature vectors. These image vectors are trained with six different ML algorithms such as KNN, DT, NB, LR, SVM, and RF to produce the classification model. The Random Forest with FFOA attained the best performance scores of 99.54% accuracy, 99.41% recall, 99.50% precision, and 99.46% F1-score when compared with other classifiers for COVID-19 detection in the X-rays dataset. The Random Forest with FFOA classifier attained the best performance scores of 97.37% accuracy, 97.59% recall, 98.52% precision, and 98.06% F1-score when compared with other classifiers. In future work, further planning to implement image augmentation techniques to avoid the problems of imbalance of images in the dataset and to implement Alex Net architecture with extracted features for COVID-19 detection.

References

[1]. Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest CT for COVID-19: Comparison to RT-PCR, Radiology 296 (2) (2020) E115–E117

[2]. V. Shah, R. Keniya, A. Shridharani, M. Punjabi, J. Shah, N. Mehendale, Diagnosis of COVID-19 using CT scan images and deep learning techniques, Emerg. Radiol. (2021) 1–9

[3]. P.K. Chaudhary, R.B. Pachori, FBSED based automatic diagnosis of COVID-19 using X-ray and CT images, Comput. Biol. Med. 134 (2021) 104454

[4]. P. Gaur, V. Malaviya, A. Gupta, G. Bhatia, R.B. Pachori, D. Sharma, COVID-19 disease identification from chest CT images using empirical wavelet transformation and transfer learning, Biomed. Signal Process. Control 71 (2022) 103076

[5]. Navid Ghassemi, Afshin Shoeibi, Marjane Khodatars, Jonathan Heras, Alireza Rahimi, Assef Zare, Yu-Dong Zhang , Ram Bilas Pachori. Manuel Gorriz, Automatic diagnosis of COVID-19 from CT images using CycleGAN and transfer learning, Applied Soft Computing 144 (2023) 110511

[6]. H. Mary Shyni, E. Chitra, A comparative study of X-ray and CT images in COVID-19 detection using image processing and deep learning techniques, Computer Methods and Programs in Biomedicine Update 2 (2022) 100054

[7]. G. U. Nneji, J. Cai, J. Deng, H. N. Monday, E. C. James, and C. C. Ukwuoma, "Multi-Channel Based Image Processing Scheme for Pneumonia Identification," Jan. 27, 2022, Multidisciplinary Digital Publishing Institute. doi: 10.3390/diagnostics12020325.

[8]. Sohaib Asif, Ming Zhao, Fengxiao Tang, Yusen Zhu, A deep learning-based framework for detecting COVID-19 patients using chest X-rays, Multimedia Systems (2022) 28:1495–1513

[9]. Sara Hosseinzadeh Kassania, Peyman Hosseinzadeh Kassasnib, Michal J. Wesolowskic, Kevin A. Schneidera, Ralph Deters, Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach, arXiv:2004.10641v1 [eess.IV] 22 Apr 2020

[10]. Karrar Hameed Abdulkareem, Salama A. Mostafa, Zainab N. Al-Qudsy, Mazin Abed Mohammed, Alaa S. Al-Waisy, Seifedine Kadry, Jinseok Lee, and Yunyoung Nam, Automated System for Identifying COVID-19 Infections in Computed Tomography Images Using Deep Learning Models, Journal of Healthcare Engineering Volume 2022, Article ID 5329014

[11]. Mundher Mohammed Taresh, Ningbo Zhu, Talal Ahmed Ali Ali, Asaad Shakir Hameed, and Modhi Lafta Mutar, Transfer Learning to Detect COVID-19 Automatically from X-Ray Images Using Convolutional Neural Networks, Hindawi International Journal of Biomedical Imaging Volume 2021, Article ID 8828404

[12]. S.K. Towfek, Ehsaneh khodadadi, Fatma M. Talaat, Automated Detection and Segmentation of COVID-19 Infection using Machine Learning, Journal of Artificial Intelligence and Metaheuristics (JAIM) Vol. 03, No. 02, PP. 28-37, 2023

[13]. Nabila Mansouri, Khalid Sultan, Aakash Ahmad, Ibrahim Alseadoon and Adal Alkhalil, A Deep Learning Framework for COVID-19 Diagnosis fromComputed Tomography, Intelligent Automation & Soft Computing DOI:10.32604/iasc.2022.025046

[14]. Zhang F. Application of machine learning in CT images and X-rays of COVID-19 pneumonia. Medicine 2021; 100:36 (e26855).

[15]. P. V. Naresh, R. Visalakshi, B. Satyanarayana, Automatic Detection of Covid-19 from Chest X-Rays using Weighted Average Ensemble Framework, International Journal of Intelligent Systems and Applications in Engineering IJISAE, 2024, 12(2s), 608–614

[16]. Hossein Mohammad-Rahimi, Mohadeseh Nadimi, Azadeh Ghalyanchi-Langeroudi, Mohammad Taheri and Soudeh Ghafouri-Fard, Application of Machine Learning in Diagnosis of COVID-19 Through X-Ray and CT Images: A Scoping Review, Frontiers in Cardiovascular Medicine, March 2021, Volume 8, Article 638011

[17]. Ahmad Mozaffer Karim, Hilal Kaya, Veysel Alcan, Baha Sen and Ismail Alihan Hadimlioglu, New Optimized Deep Learning Application for COVID-19 Detection in Chest X-ray Images, Symmetry 2022, 14, 1003.

[18]. M. A. Khan et al., “COVID-19 Classification from Chest X-Ray Images: A Framework of Deep Explainable Artificial Intelligence,” Jul. 14, 2022, Hindawi Publishing Corporation. doi: 10.1155/2022/4254631.

[19]. Yassine Meraihi, Asma Benmessaoud Gabis, Seyedali Mirjalili, Amar Ramdane-Cherif, Fawaz E. Alsaadi, Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey, SN Computer Science (2022) 3:286

[20]. Vikas Kumar, Arpit Gupta, Barenya Bikash Hazarika and Deepak Gupta, Automatic diagnosis of COVID-19 from Chest Xray images using transfer learning-based deep features and machine learning models, March 2022, https://doi.org/10.21203/rs.3.rs-1410155/v1

[21]. https://data.mendeley.com/datasets/8h65ywd2jr/3

[22]. 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.

[23]. 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.

[24]. Humeau-Heurtier, A.: Texture Feature Extraction Methods: A Survey. 7, 26 (2019) https://doi.org/10.1109/AC-CESS.2018.2890743.

[25]. T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[26]. J. R. Quinlan, “Induction of decision trees,” Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[27]. P. Langley,W. Iba, and K. Thompson, “An analysis of bayesian classifiers,” in AAAI, vol. 90, pp. 223–228, 1992.

[28]. Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), 221

[29]. B. Schölkopf and C. J. Burges, Advances in kernel methods: support vector learning. MIT press, 1999.

[30]. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[31]. Rajinikanth, Venkatesan et al. “Firefly-Algorithm Supported Scheme to Detect COVID-19 Lesion in Lung CT Scan Images using Shannon Entropy and Markov-Random-Field.” *arXiv: Image and Video Processing* (2020): n. pag.