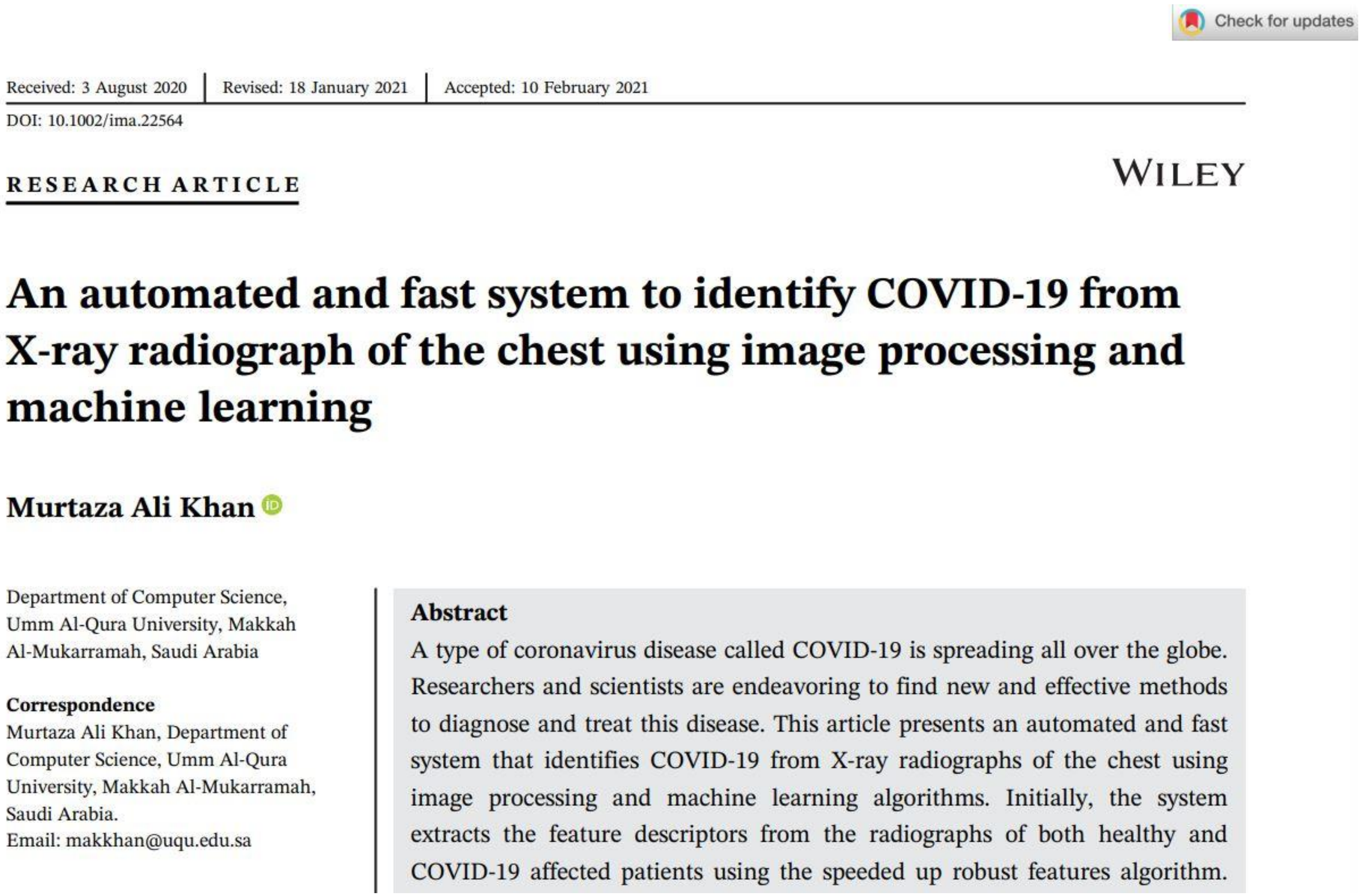
Machine Learning Project

Amin Jomepour – Sina Soltani

دانشگاه شیراز

Article Info



Received: 3 August 2020 | Revised: 18 January 2021 | Accepted: 10 February 2021

DOI: 10.1002/ima.22564

RESEARCH ARTICLE

WILEY

An automated and fast system to identify COVID-19 from X-ray radiograph of the chest using image processing and machine learning

Murtaza Ali Khan

Department of Computer Science,
Umm Al-Qura University, Makkah
Al-Mukarramah, Saudi Arabia

Correspondence
Murtaza Ali Khan, Department of
Computer Science, Umm Al-Qura
University, Makkah Al-Mukarramah,
Saudi Arabia.
Email: makkhan@uqu.edu.sa

Abstract

A type of coronavirus disease called COVID-19 is spreading all over the globe. Researchers and scientists are endeavoring to find new and effective methods to diagnose and treat this disease. This article presents an automated and fast system that identifies COVID-19 from X-ray radiographs of the chest using image processing and machine learning algorithms. Initially, the system extracts the feature descriptors from the radiographs of both healthy and COVID-19 affected patients using the speeded up robust features algorithm.

Research Goal

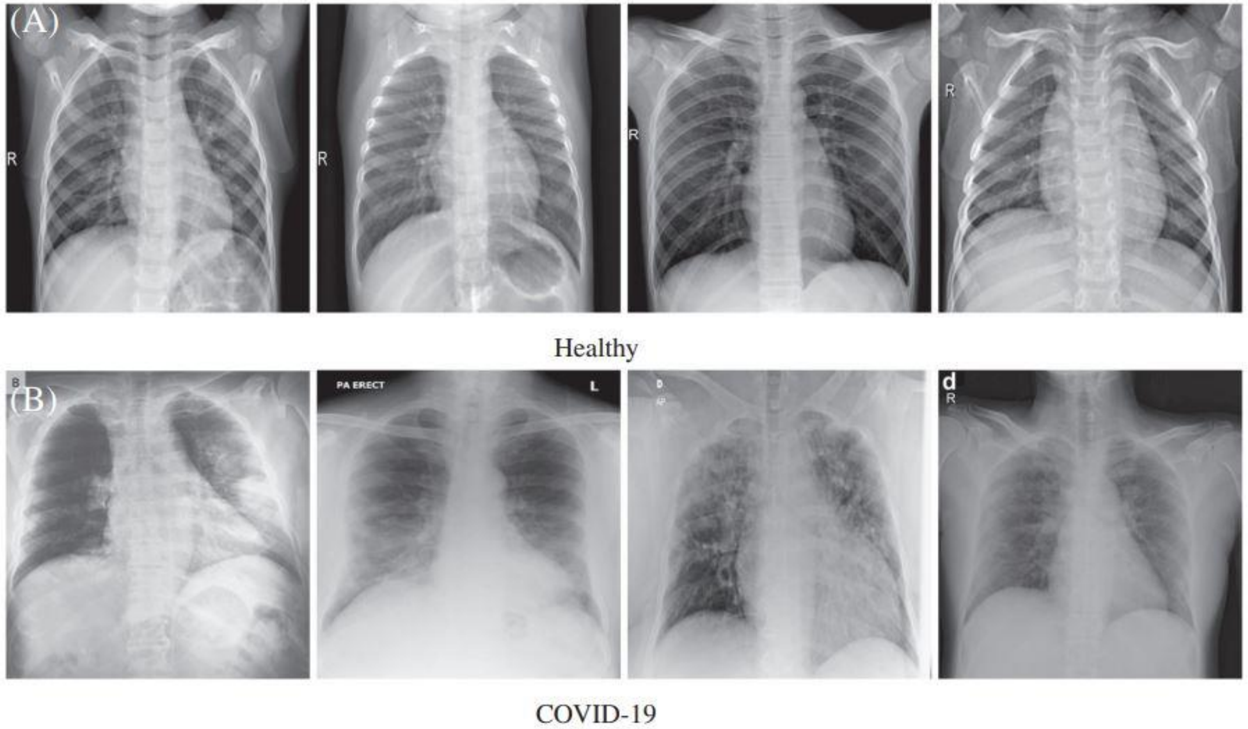
- This article presents an automated and fast system that identifies COVID-19 from X-ray radiographs of the chest using image processing and machine learning algorithms.

Intro

- Radiography of the chest/lungs uses an X-ray to help diagnose the cause of several lungs disorders including COVID-19.
- The underlying assumption of the study is that COVID-19 affects the lungs most that exhibit certain characteristics (features) in the lungs tissues.
- Image processing and machine learning algorithms can detect the COVID-19 using the X-ray radiographs of the chest.

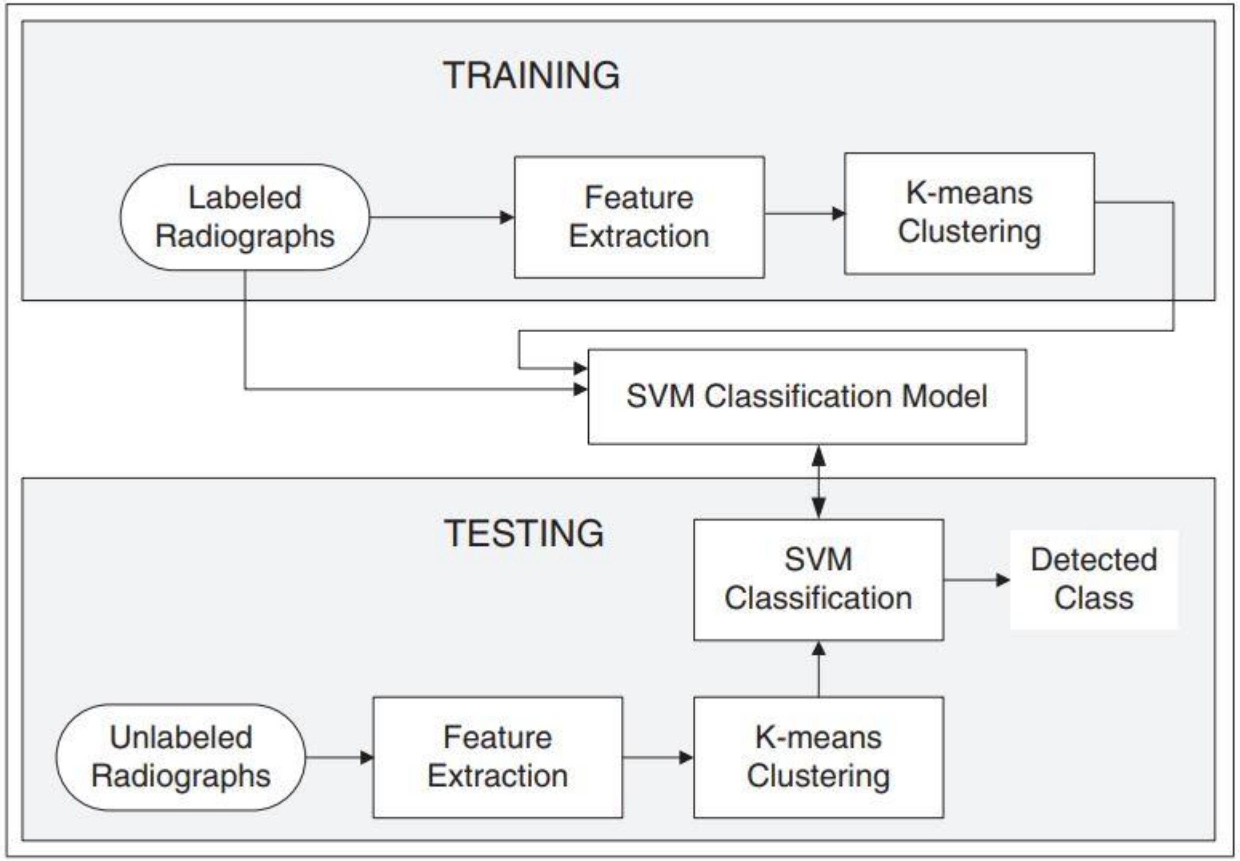
Dataset

The study used the dataset of 340 X-ray radiographs, 170 images of each Healthy and Positive COVID-19 class.



System Architecture and Methodology

1. Extraction of feature descriptors.
2. Clustering of feature descriptors.
3. Classification of radiographs.

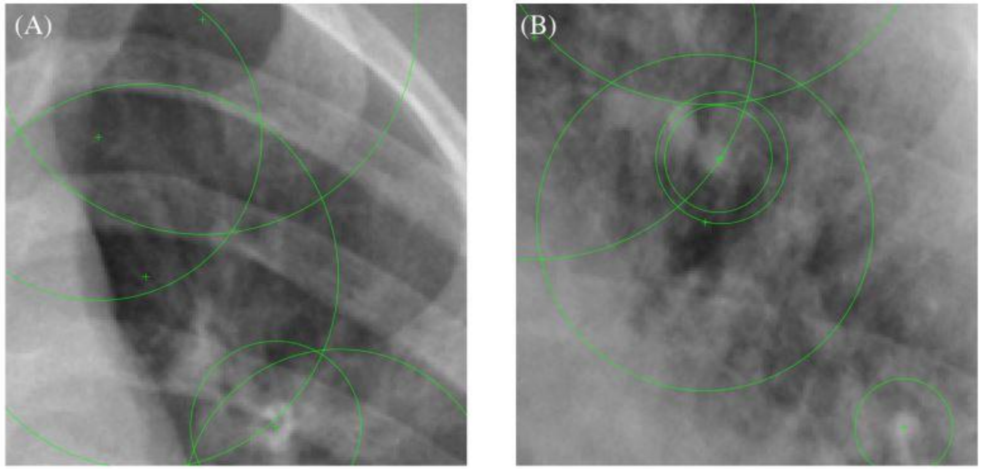


Step 1: Extraction of feature descriptors

- A feature descriptor is a set of values that describes the image patch around the point of interest. Local features extractors detects patterns or structures such as a corner, curvature, or edges in a grayscale image via local minima/maxima of some function.
- **SURF** offers a computationally efficient approximation of the second-order Gaussian derivatives via a set of integral images. For the detection of feature points, instead of relying on perfect Gaussian derivatives, the computation is based on simple two-dimensional box filters.
- SURF employs a scale-invariant blob detector that relies on the determinant of the Hessian matrix for selection of scale selection.

SURF algorithm output

- The SURF algorithm returns the following information for a feature descriptor:
  1. Spatial coordinates, that is, (x, y) of the feature point.
  2. The scale at which the feature point is detected.
  3. Strength of the feature point.
  4. Sign of the Laplacian operator. This value must be an integer −1, 0, or 1.
  5. The orientation of the feature point in radian.



Healthy

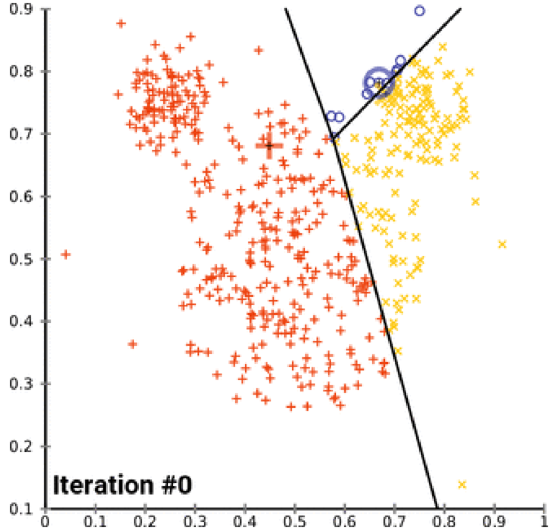
COVID-19

Step 2: Clustering of feature descriptors

- The system detects a very large number of feature descriptors from chest radiographs. This big feature descriptor-vocabulary needs to be reduced to some manageable size.
- They group the descriptors obtained from all the radiographs into 500 clusters using the **K-means algorithm**.
- The clusters are mutually exclusive and smaller in number compared to the number of descriptors. The center of each cluster represents a visual-word. The centers of all the clusters of all the radiographs build the visual-vocabulary.

K-means Algorithm

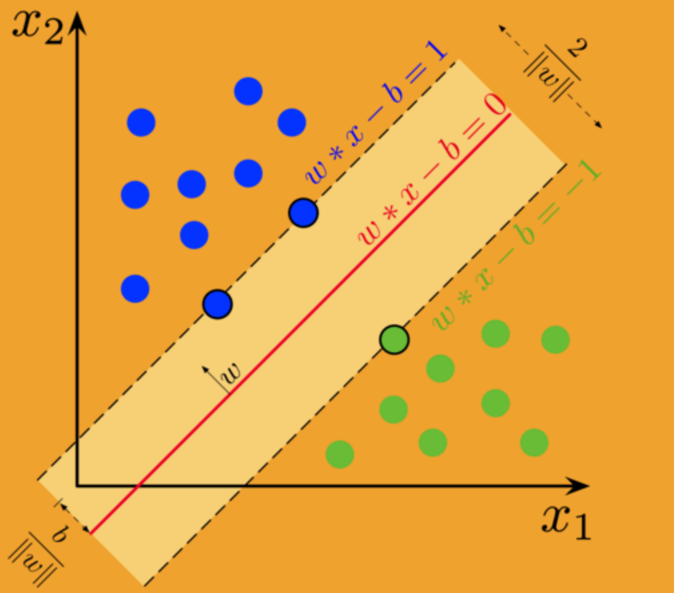
1. Randomly choose centroid of each of the K clusters.
2. Allocate ith pixel of the radiograph image to a cluster that minimizes the Euclidean distance between the ith pixel and the centroid.
3. Recompute the centroid of each cluster by averaging all the pixels in the cluster.
4. Repeat steps 2 and 3 until convergence is achieved, that is, any pixel does not change its centroid.



Step 3: Classification of images

- the **SVM classifier** is trained for two labeled (known) radiograph sets of Healthy and Positive COVID-19.
- During the testing, the system is inquired to find the class of an unlabeled (unknown) test radiograph. The SVM classifier compares the feature descriptors of the test radiograph with the visual vocabulary of the classifier. The class of the test radiograph is predicted based on the optimal match.

SVM Classifier



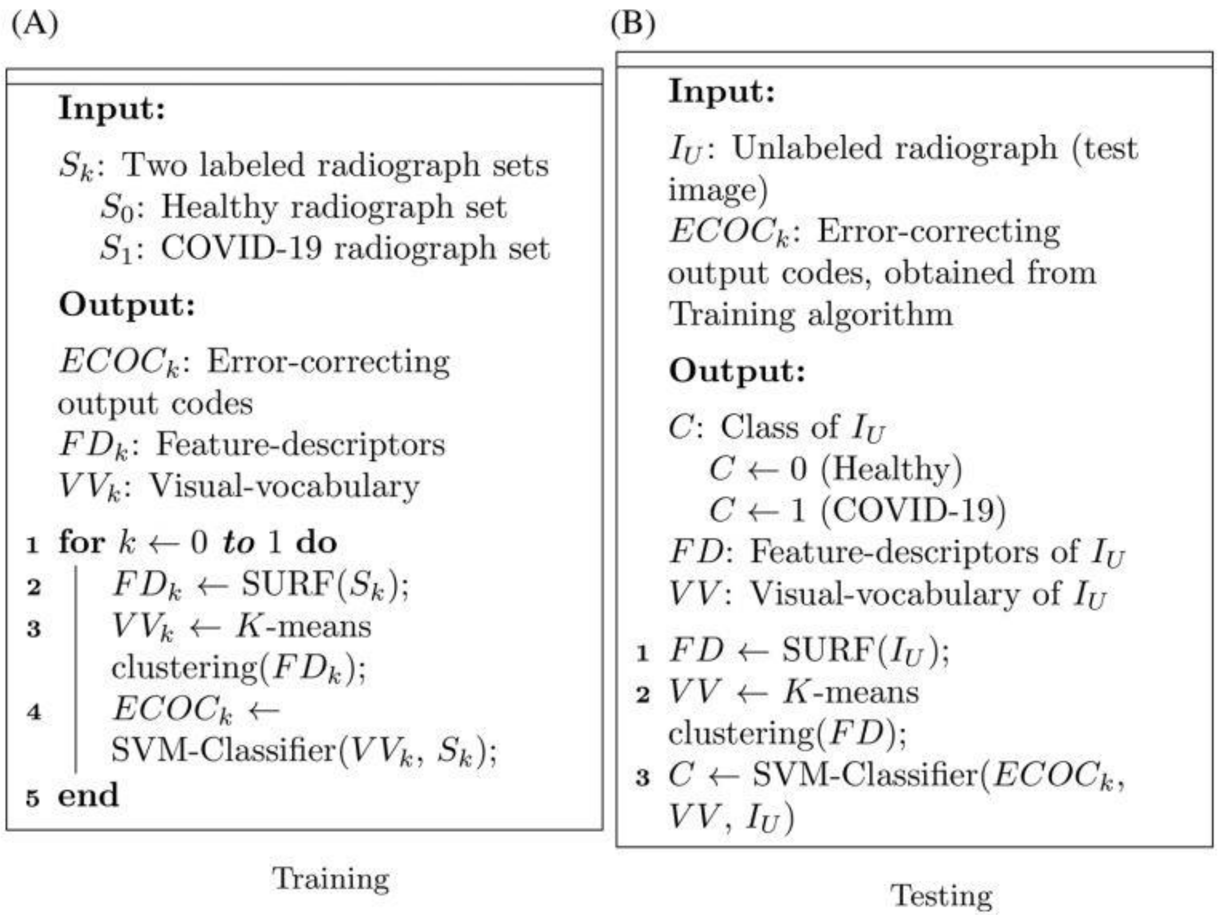
- A binary SVM constructs an optimal hyperplane in high-dimensional space that separates data into two classes labeled as −1 and 1.
- The optimal hyperplane is one that maximizes the margin-width between two classes.
- The equation of hyperplane H can be written as follows.

$$H : w.x - b = 0,$$

- where $b \in R$ is called the bias and $w \in R^n$ is referred to as the weight vector.

$$w.x_i - b > 0, y_i = 1,$$

$$w.x_i - b \leq 0, y_i = -1.$$



(A)

```
Input:
    $S_k$: Two labeled radiograph sets
        $S_0$: Healthy radiograph set
        $S_1$: COVID-19 radiograph set
Output:
    $ECOC_k$: Error-correcting output codes
    $FD_k$: Feature-descriptors
    $VV_k$: Visual-vocabulary
for $k \leftarrow 0$ to 1 do
    $FD_k \leftarrow$ SURF($S_k$);
    $VV_k \leftarrow$ $K$-means clustering($FD_k$);
    $ECOC_k \leftarrow$ SVM-Classifier($VV_k$, $S_k$);
end
```

Training

(B)

```
Input:
    $I_U$: Unlabeled radiograph (test image)
    $ECOC_k$: Error-correcting output codes, obtained from Training algorithm
Output:
    $C$: Class of $I_U$
        $C \leftarrow 0$ (Healthy)
        $C \leftarrow 1$ (COVID-19)
    $FD$: Feature-descriptors of $I_U$
    $VV$: Visual-vocabulary of $I_U$
$FD \leftarrow$ SURF($I_U$);
$VV \leftarrow$ $K$-means clustering($FD$);
$C \leftarrow$ SVM-Classifier($ECOC_k$, $VV$, $I_U$)
```

Testing

FIGURE 3 Algorithms

Evaluation Matrices

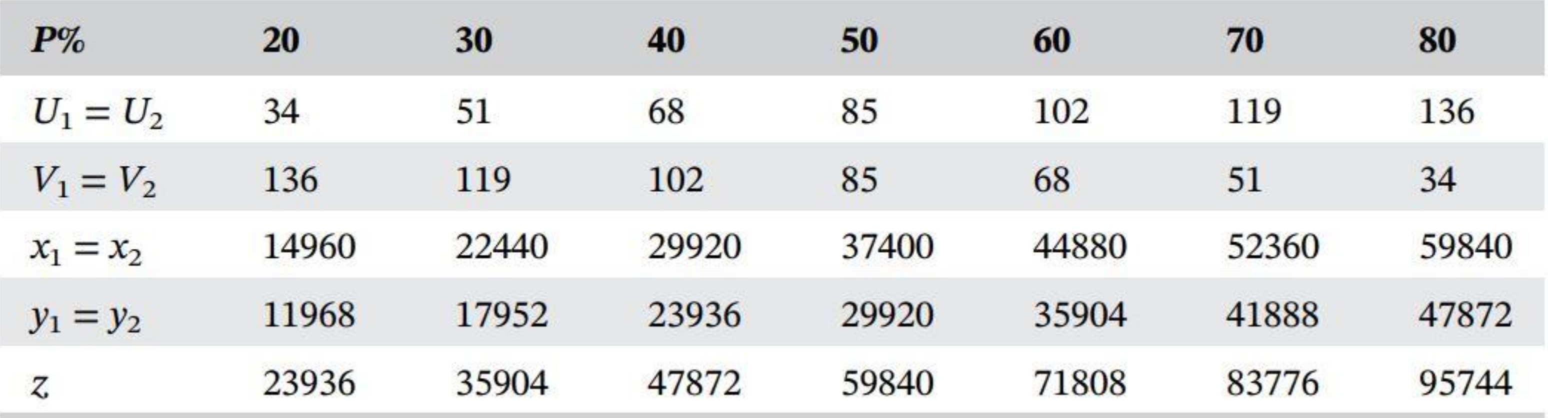
- Accuracy is the ratio of correctly classified radiographs to the total number of radiographs classified.
- Mathematically, the following is the equation of percentage accuracy. (A):

$$A = \frac{(TN + TP)}{(TN + TP + FN + FP)} \times 100.$$

- Where True negative (TN), True Positive (TP), False negative (FN), False Positive (FP).

Experiments

- The datasets of 340 X-ray radiographs used in the experiments. We resize them to 360 × 320(width × height). Resizing the images also speed up both the training and testing algorithms.
- Let there are total N radiographs in the dataset such that N1 radiographs belong to the Healthy class, while N2 radiographs are of positive COVID-19 cases.
- The z features are passed to the K-means clustering algorithm to create 500 mutually exclusive clusters.



P%	20	30	40	50	60	70	80
$U_1 = U_2$	34	51	68	85	102	119	136
$V_1 = V_2$	136	119	102	85	68	51	34
$x_1 = x_2$	14960	22440	29920	37400	44880	52360	59840
$y_1 = y_2$	11968	17952	23936	29920	35904	41888	47872
z	23936	35904	47872	59840	71808	83776	95744

Classification using Deep-Learning based CNN

- Although this work's main contribution is to build a SVM-based system to identify positive and negative COVID-19 cases, They also implemented deep learning based CNN to classify COVID-19 images for comparison.
- Following layers are used to build the CNN:
  1. The input-layer feed radiographs to a network and applies data normalization
  2. The convolutional-layer applies convolutional filters to the radiographs by moving the filters along the rows and columns of images and computing the dot product of the weights and the input and then adding a bias term.
  3. The Batch-normalization-layer first normalizes each channel's activations by subtracting the mini-batch mean and dividing by the mini-batch SD. Then, the layer shifts the input by learning offset β and scales it by learning scaling factor γ.
  4. The rectified linear unit layer set all the negative input elements to zero
  5. The fully connected-layer multiplies the input by a weight matrix and then adds a bias vector.
  6. The softmax-layer applies a softmax function to the input.
  7. The classification-layer computes the cross-entropy loss for multiclass classification problems with mutually exclusive classes.

RESULTS

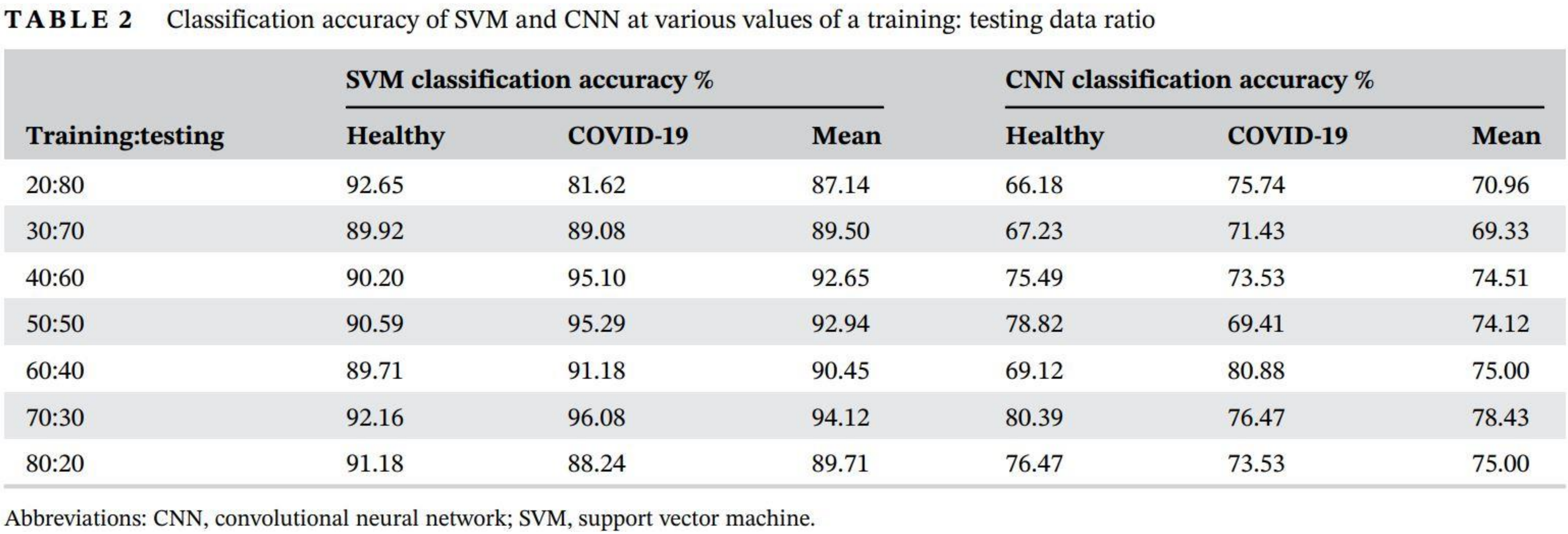
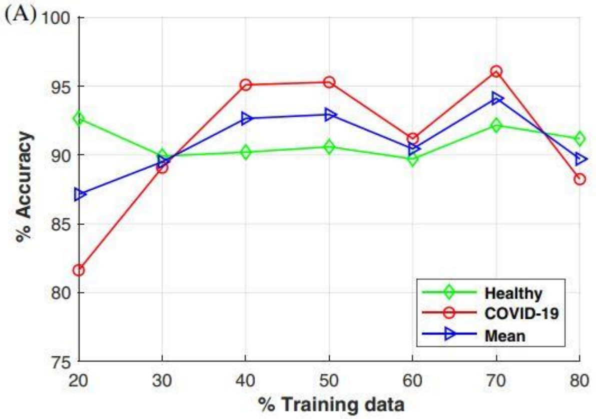


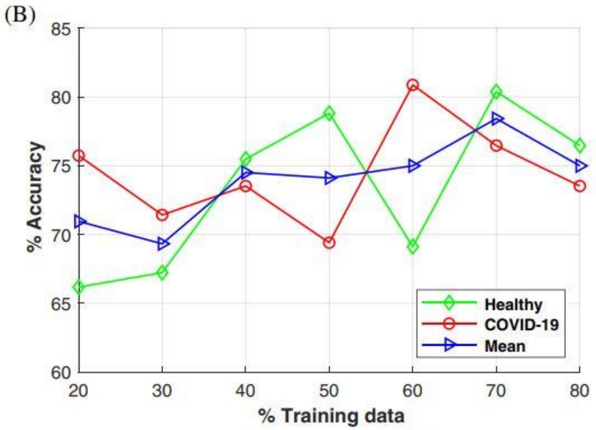
TABLE 2 Classification accuracy of SVM and CNN at various values of a training: testing data ratio

	SVM classification accuracy %			CNN classification accuracy %		
Training:testing	**Healthy**	**COVID-19**	**Mean**	**Healthy**	**COVID-19**	**Mean**
20:80	92.65	81.62	87.14	66.18	75.74	70.96
30:70	89.92	89.08	89.50	67.23	71.43	69.33
40:60	90.20	95.10	92.65	75.49	73.53	74.51
50:50	90.59	95.29	92.94	78.82	69.41	74.12
60:40	89.71	91.18	90.45	69.12	80.88	75.00
70:30	92.16	96.08	94.12	80.39	76.47	78.43
80:20	91.18	88.24	89.71	76.47	73.53	75.00

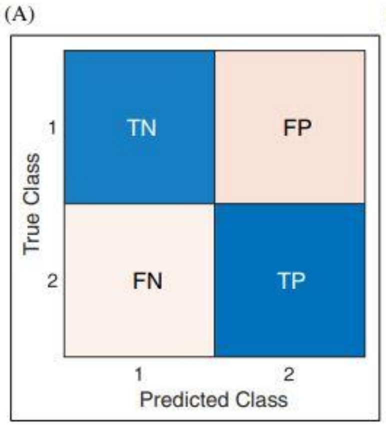
Abbreviations: CNN, convolutional neural network; SVM, support vector machine.



Accuracy of SVM

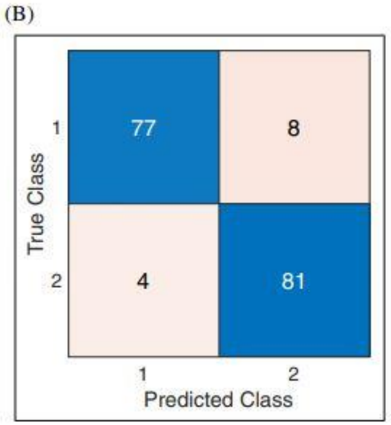


Accuracy of CNN


(A)
True Class
Predicted Class

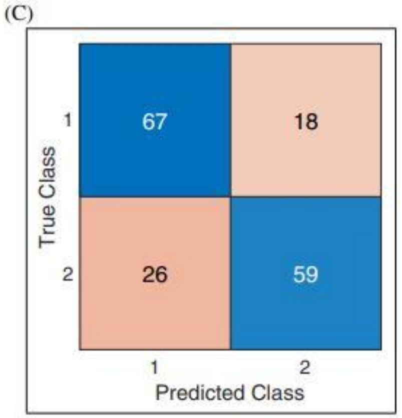
	1	2
1	TN	FP
2	FN	TP

Confusion Matrix (CM)


(B)
True Class
Predicted Class

	1	2
1	77	8
2	4	81

CM of SVM


(C)
True Class
Predicted Class

	1	2
1	67	18
2	26	59

CM of CNN