Original Article

Role of Optimal Features Selection with Machine Learning Algorithms for Chest X-ray Image Analysis

Mohini Manav[1,2], Monika Goyal[1], Anuj Kumar[2]

[1]Department of Physics, GLA University, Mathura, [2]Department of Radiotherapy, S N Medical College, Agra, Uttar Pradesh, India

Abstract

Introduction: The objective of the present study is to classify chest X-ray (CXR) images into COVID-positive and normal categories with the optimal number of features extracted from the images. The successful optimal feature selection algorithm that can represent images and the classification algorithm with good classification ability has been determined as the result of experiments. **Materials and Methods:** This study presented a framework for the automatic detection of COVID-19 from the CXR images. To enhance small details, textures, and contrast of the images, contrast limited adaptive histogram equalization was used. Features were extracted from the first-order statistics, Gray-Level Co-occurrence Matrix, Gray-Level Run Length Matrix, local binary pattern, Law's Texture Energy Measures, Discrete Wavelet Transform, and Zernikes' Moments using an image feature extraction tool "pyFeats. For the feature selection, three nature-inspired optimization algorithms, Grey Wolf Optimization, Particle Swarm Optimization (PSO), and Genetic Algorithm, were used. For classification, Random Forest classifier, K-Nearest Neighbour classifier, support vector machine (SVM) classifier, and light gradient boosting model classifier were used. **Results and Discussion:** For all the feature selection methods, the SVM classifier gives the most accurate and precise result compared to other classification models. Furthermore, in feature selection methods, PSO gives the best result as compared to other methods for feature selection. Using the combination of the SVM classifier with the PSO method, it was observed that the accuracy, precision, recall, and F1-score were 100%. **Conclusion:** The result of the study indicates that with optimal features with the best choice of the classifier algorithm, the most accurate computer-aided diagnosis of CXR can be achieved. The approach presented in this study with optimal features may be utilized as a complementary tool to assist the radiologist in the early diagnosis of disease and making a more accurate decision.

Keywords: Artificial intelligence, chest X-ray, feature selection, image classification, machine learning

Received on: 23-11-2022 **Review completed on: 26-04-2023** **Accepted on: 06-05-2023** **Published on: 29-06-2023**

Introduction

Coronavirus disease (COVID-19), caused by the SARS-CoV-2 virus, is one of the deadliest and most contagious diseases of this decade.[1] The COVID-19 pandemic can be considered severe due to the high rate of transmission and lethality.[2] Lack of any prior knowledge about this disease overwhelmed hospitals with shortages of detection tools and medical supplies, making the battle against COVID-19 onerous.[3] Most infected people experience mild-to-moderate respiratory symptoms.[4] Effective screening of the infected people was a critical step. Many techniques, such as real-time polymerase chain reaction, Truenat screening, cartridge-based nucleic acid amplification test, rapid antibody, and rapid antigen test techniques, were used for the detection of COVID-19.[5,6]

The chest X-ray (CXR) is a useful, noninvasive clinical adjunct that aids in the initial diagnosis of a variety of pulmonary disorders.[7] In the early diagnosis of COVID-19 disease, a vital role was played by the imaging techniques like CXR imaging. Some studies revealed distinct visual characteristics such as ground-glass opacities and patchy reticular opacities.[8] Ground-glass opacities, patchy reticular opacities, can be detected on CXR images but not precisely as much in computed tomography (CT) images. CT is a more precise approach for imaging the chest, but it has not replaced CXR as the major imaging test due to the additional time, cost, and radiation exposure involved with CT. Less ionizing radiation, quick data collection, accessibility in intensive care units, and mobility

Access this article online

Quick Response Code:

Website: www.jmp.org.in

DOI: 10.4103/jmp.jmp_104_22

Address for correspondence: Dr. Anuj Kumar, Department of Radiotherapy, S N Medical College, Agra - 282 002, Uttar Pradesh, India.
E-mail: toaktyagi@gmail.com

How to cite this article: Manav M, Goyal M, Kumar A. Role of optimal features selection with machine learning algorithms for chest X-ray image analysis. J Med Phys 2023;48:195-203.

are a few of the advantages that CXR has over CT. However, it is difficult and requires subject expert knowledge to manually identify these minor visual characteristics on CXR images.[9]

With the development in technology, computer-aided diagnosis systems are growing in popularity since they reduce the cost of setting up a medical laboratory, hence expanding access to quality health care.[10] Artificial intelligence has offered a new glimmer of hope for assisting clinicians in making more accurate imaging diagnoses and decreasing their workload.[11,12]

Identifying the most significant risk factor associated with an illness is crucial in medical diagnosis. Quantitative data values or pixel intensities that provide rich, meaningful information about the pixels of an image in terms of local and/or global variation constitute an image's features.[13] Image features can be handcraft features and nonhandcraft features or learned features.[14] Handcraft features are manually engineered features and learning features are the features that are automatically extracted by the algorithm, like in deep learning.

Unfortunately, sometimes, deep learning models may suffer from overfitting problems, cause high bias because they extract unknown and abstract features, and need high-dimensional datasets to obtain higher performance.[15-17] To overcome such problems, some researchers used pretrained transfer learning models to take advantage of the potential of deep learning techniques.[18] For a specific problem, more meaningful and more known features designed can be extracted manually using handcrafted features. Furthermore, such techniques do not require a lot of data.[19,20]

Multiple features-based classifications have the advantage of achieving good classification accuracy. With the selection of the most effective image features, all noise, redundant, and interrelated features can be removed.[21] However, if too many features are derived from a limited training dataset, then due to overfitting, the robustness of the classification model will decrease on data other than the data used for training the model.[22] Hence, the selection of a limited number of features is necessary to balance the accuracy and robust classification efficiency of the model. There are several methods available for feature selection like forward selection, backward elimination, recursive feature elimination (RFE), linear discriminant analysis, etc.[23-25] Chandra *et al*. used 8196 features from the CXRs, which are eight first-order statistical features (FOSF), 88 Gray-Level Co-Occurrence Matrix (GLCM), and 8100 histograms of oriented gradients and they selected the features Using Binary Gray wolf optimization.[26] Öztürk *et al*. used GLCM, local binary GLCM, Gray-level run length matrix (GLRLM), and segmentation-based fractal texture analysis features and to identify the most prominent features principal component analysis (PCA) was used.[27] Bhargava *et al*. presented a comparative study that compared the classification performance of four machine learning models, i.e., PCA, K-Nearest Neighbour (KNN), Sparse Representation Classifier, artificial neural network (ANN), and SVM, for the different features extracted from the segmented CT and CXR images.[28] Kumar *et al*. proposed the notion of Pearson Correlation Coefficient along with variance thresholding to optimally reduce the feature space of extracted features from the conventional deep learning architectures from the CXR images and used these features to classify the images into different categories.[29] Sethy *et al*. presented a 2-class study using the SVM algorithm with the features obtained from 11 different well-known convolutional neural network (CNN) models with a very little data set. In their second study, they performed a 3-class study with increased data and determined the classification performance with SVM using the feature maps obtained from 13 CNN models.[30]

The objective of the present study is to classify the CXR images into a COVID-positive and normal categories with the optimal number of features extracted from the images. The successful optimal feature selection algorithm that can represent images and the classification algorithm with good classification ability has been determined as the result of experiments.

Materials and Methods

Dataset used

The presented work used publicly available data, the COVID-QU-Ex dataset from the Kaggle database, which was compiled by researchers from Qatar University.[31-35] In the present study, out of 33,920 images, 400 CXR images were randomly selected with their ground truth to study only the lung part. All images were in portable network graphics format. The ability to segment lungs from surrounding structures significantly reduces the execution time and boosts the effectiveness of nodule identification.[36] The given lungs-only masks from the database were used to segment the lungs from the X-ray images.

Data preprocessing

The objective of the data pre-processing stage is to prepare the data for use in the prediction model. Typically, data are disorganized and derived from various sources with varying sizes and resolutions. To reduce the complexity and increase the accuracy of the prediction model, cleaning up the data is crucial. In the first step, all the images were resized to 256 × 256. After that, to enhance small details, textures, and contrast of the images, contrast limited adaptive histogram equalization, which is an adaptive contrast histogram equalization method, was used.[37]

Features extraction

In any pattern classification system, feature extraction is a crucial step. Features were extracted from the first-order statistics (FoS), GLCM, GLRLM, local binary pattern (LBP), Law's Texture Energy Measures (LTEM), Discrete Wavelet Transform (DWT), and Zernikes' Moments (ZM) using an image feature extraction tool "pyFeats."[38-46]

First-order statistical features

These features were extracted from the histogram of the image. These features included mean, median, skewness, kurtosis, etc.

Gray-level co-occurrence matrix

It is a second-order statistical feature that provides information about texture. GLCM examines the spatial relationship among pixels and defines the frequency of a combination of pixels in a given direction and distance. For texture analysis, several features like angular second moment, contrast, entropy, correlation, etc., can be calculated from the GLCM.

Gray-level run length matrix

It is a 2-D matrix of the component in which each component provides information about the total number of occurrences of the run having a length n of the grey level in a given direction. In GLRLM, the run length, or the number of the neighboring grey levels in a particular direction, is the frequency of the particular run in the image. Features such as short-run emphasis, long-run emphasis, run percentage, low gray level run emphasis, high gray level run emphasis, etc., can be calculated from GLRLM.

Local binary pattern

It is a texture descriptor that describes the local texture pattern of the image and is used for the property of high discriminative power. It assigns a binary number to the pixel by comparing the gray level with the neighboring after labeling. Energy and entropy of the LBP image, constructed over different scales, are used as feature descriptors.

Law's texture energy measures

A texture-energy-based approach developed by Law, measures the amount of variation within a fixed-size window. It uses a set of convolutional kernels to compute the texture energy. Features such as texture energy from different kernels, etc., can be calculated from the resulting image.

Discrete wavelet transform

DWT provides information on the time and frequency of the signal simultaneously. The different information of the main signal, such as high-frequency or low-frequency segments, can be extracted with the help of wavelet decomposition. By the application of DWT, the image is divided into four sub-bands, i.e., low-frequency components in horizontal and vertical directions (cA), the low-frequency component in the horizontal, and high-frequency component in the vertical direction (cV), the high-frequency component in the horizontal and low-frequency component in the vertical direction (cH) and high-frequency components in horizontal and vertical directions (cD). cA, cV, cH, and cD can also be represented as LL, LH, HL, and HH, respectively.

Zernikes' moments

In image shape description and content-based image retrieval, the moments, such as geometric moments, centric moments, and orthogonal invariance moments, have been used. The most commonly used technique in image shape feature extraction and description, i.e., Zernike moment, which is one kind of the orthogonal invariance moment and its kernel is a set of Zernike complete orthogonal polynomials defined over the interior of the unit disc in the polar coordinates space. An image moment is a certain particular weighted average of the image pixels' intensities in image processing, computer vision, and related fields or a function of such moments, usually chosen to have some attractive property or interpretation.

Features selection

The features extracted differed in their range, so to make them on the same scale, standardization was performed.[47] To reduce the computation times and to take up less storage space, dimensionality reduction and feature selection can be so helpful. With the help of feature selection methods, we can remove noisy, redundant data and optimize the efficiency of the classification model simultaneously.[48]

For the feature selection, three nature-inspired optimization algorithms were used. These nature-inspired algorithms were Grey Wolf Optimization (GWO) algorithm, Particle Swarm Optimization (PSO) algorithm, and Genetic Algorithm (GA).[49-51] For feature selection using the GWO, PSO, and GA python library Zoofs was used.[52]

Grey wolf optimization

It is a population-based meta-heuristic algorithm. It stimulates the leadership and predation behavior of the grey wolves. For simulating the leadership hierarchy, four types of grey wolves such as alpha, beta, gamma, and omega are employed. The three main steps of hunting, searching for prey (exploration), encircling prey, and attacking prey, are implemented in this algorithm. In the present study, the value of the parameters, i.e., size of the population and number of iterations used for GWO, were 40 and 20, respectively.

Particle swarm optimization

It is another metaheuristic optimization algorithm. It is inspired by swarm behavior such as a school of fish or a swarm of birds. It is one of the bio-inspired algorithms and differs from other optimization algorithms as it is not dependent on the gradient or any differential form of the objective. Due to the easy encoding of features, global search facility, being reasonable computationally, fewer parameters, and easier implementation, PSO is a suitable algorithm for feature selection. In the present study, the parameters used for PSO were the size of the population, number of the iteration, first acceleration coefficient of the particle swarm, second acceleration coefficient of the particle swarm, and weight parameter. The values of the above parameters used in the present study were 40, 20, 2, 2, and 0.9, respectively.

Genetic algorithm

It is an adaptive heuristic search algorithm based on the principles of genetics and natural selection. In GAs, there is a population or pool of possible solutions to the given problem. These solutions are then subjected to recombination and mutation (similar to natural genetics), resulting in the birth of new offspring, and the process is repeated for multiple generations. Each individual (or candidate solution) is assigned a fitness value, and the fitter people are given a greater opportunity to mate and produce more "fit" individuals. It

optimizes both continuous and discrete functions and also multi-objective problems. It is very useful when the search space is very large and there are a large number of parameters involved. In the present study, the parameters used for GA were the size of the population, number of the iteration, selective pressure, elitism, and mutation rate. The values of the above parameters used in the present study were 40, 20, 2, 2, and 0.05, respectively.

Image classification

For the classification of the images using the reduced features from the above-mentioned algorithms, different machine-learning models were used. These models were the Random Forest (RF) classifier, KNN classifier, SVM classifier, and light gradient boosting model (LGBM) classifier.[53-56]

Random forest

It is a supervised machine learning algorithm based on ensemble learning. It is a combination of the large number of the decision each tree in the ensemble is comprised of a bootstrap sample, which is a data sample obtained from a training set with replacement. To make the final decision, it aggregates the output from all decision trees formed on different samples and decides the outcome based on the majority voting. Figure 1 shows the working of RF.

K-nearest neighbour

It is also a nonparametric supervised machine learning algorithm. An input test data point is classified by identifying the K nearest training vectors using a suitable distance metric. The class to which the majority of these K nearest neighbors belong is then allocated to the test input data point. Figure 2 shows the working of KNN.

Support vector machine

It is a supervised machine learning algorithm, and the objective of the SVMC is to find a hyperplane in N-dimensional space that can distinctly classify the data points. The dimension of the hyperplane depends on the number of features. Data points falling on either side of the hyperplane can be attributed to different classes. It chooses the extreme points that help in creating the hyperplane. These extreme cases are called "support vectors." Figure 3 shows the working of SVM.

Light gradient boosting model

It is based on the gradient boosting framework. It uses two types of techniques which are Gradient-based one side sampling (GOSS) and Exclusive Feature bundling (EFB). In contrast to other boosting algorithms, which grow tree level-wise, it selects the leaf with maximum delta loss to grow. Limiting the maximum depth of trees can not only ensure training efficiency but also prevent overfitting. GOSS excludes the significant portion of the data part, which has small gradients and to estimate the overall information gain use the remaining data. For computation on information gain, the data instances which have large gradients play a greater

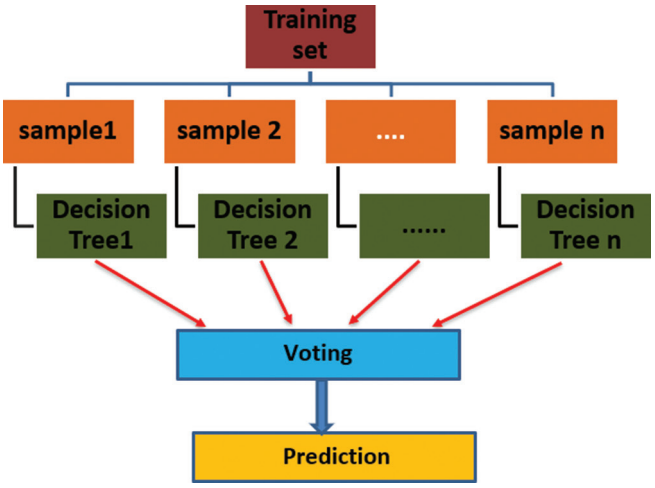

Figure 1: Working of random forest algorithm

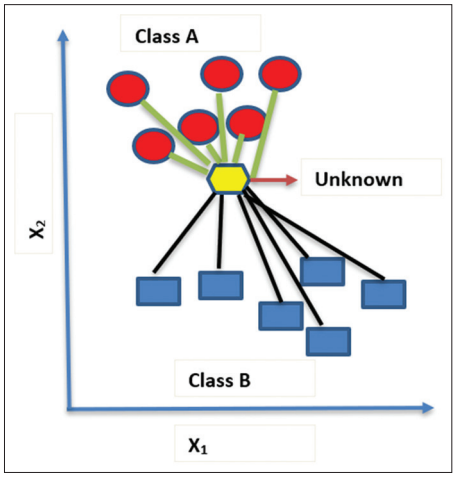

Figure 2: Working of K- nearest neighbour algorithm

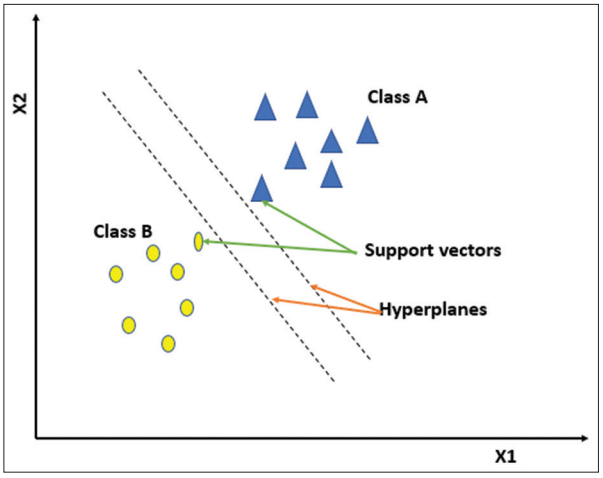

Figure 3: Working of support vector algorithm

role. When the value of information gain has a large range, this can lead to a more accurate gain estimation than uniformly random sampling with the same target sampling rate. While working with high dimensional data, there are many features

that are mutually exclusive, with the EFB technique, LGBM can safely bundle such exclusive features into a single feature to reduce the complexity.

Proposed work

In the present study, a total of 400 CXR images were taken. These images were equally divided into COVID-19 and normal classes. This dataset was further divided into training and testing in a ratio of 80:20. Figure 4 depicts the block diagram of the present study.

A total of 98 features were obtained for each image using FoS, GLCM, GLRLM, LBP, LTEM, DWT, and ZM. For optimization of the features, with each optimizing algorithm, the classification algorithm with hyper-tuned parameters was given as an objective function. For the classification models, the hyperparameters used in the study are given in Table 1.

For each classification algorithm, i.e., RF, KNN, SVM, LGBM, different numbers of optimal features were obtained using GWO, PSO, and GA. All steps, which included feature extraction, feature reduction, and classification, were implemented using Python 3.10 on Google Colab. In classification models, for hyperparameter tuning of parameters, GridSearchCV from the sklearn library was used.[57] The machine learning model is evaluated for a variety of hyperparameter values. This approach seeks for the best set of hyperparameters from a grid of hyperparameters values.

Performance evaluation

The metrics that were used in the evaluation process for the model performance were accuracy, precision, sensitivity, and F1 score. Accuracy is the ratio of the number of correct predictions to the total number of predictions. Precision is the ratio of the number of correct positive class predictions to the total positive class predictions. Sensitivity is the number of correct positive class predictions to the correct positive class predictions and the false negative class predictions. The F1 score is the weighted average of precision and sensitivity. F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The accuracy metric computes how many times a model made a correct prediction across the entire dataset. F1-score gives a better measure of the incorrectly classified cases than the accuracy metric.

Results

In the present study, the classification performance of the four machine learning algorithms (RF, KNN, SVM, LGBM) with the optimal number of features extracted from the CXR images was studied.

Feature extraction

The details of the features extracted from the Zoofs library are given in the supplementary file, with each feature name as mentioned in the library.

The different number of features extracted from the CXR images are given in Table 2.

Performances of feature selection algorithms

Using the GWO, PSO, and GA feature selection algorithms, the selected optimal number of features was selected. The GWO reduced the features in the range of 81 to 85 from

Table 1: Value of hyperparameters for different classification models

Classification model	Parameters
RF	max_features="log2," n_estimators=40
KNN	n_neighbors=6, weights="distance"
SVM	C=100, gamma=0.001
LGBM	bagging_fraction=0.5, bagging_frequency=5, feature_fraction=0.5, max_depth=10, min_data_in_leaf=110, num_leaves=1200

RF: Random Forest, KNN: K-Nearest Neighbour, SVM: Support Vector Machine, LGBM: Light Gradient Boosting Model

Table 2: Different features extracted from the chest X-ray images

Feature type	Number of features
FoS	16
GLCM	14
GLRLM	11
LBP	8
LTEM	6
DWT	18
ZM	25
Total	98

FoS: First-order statistics, GLCM: Gray-Level Co-Occurrence Matrix, GLRLM: Gray-Level Run Length Matrix, LBP: Local Binary Pattern, LTEM: Law's Texture Energy Measures, DWT: Discrete Wavelet Transform, ZM: Zernike's Moments

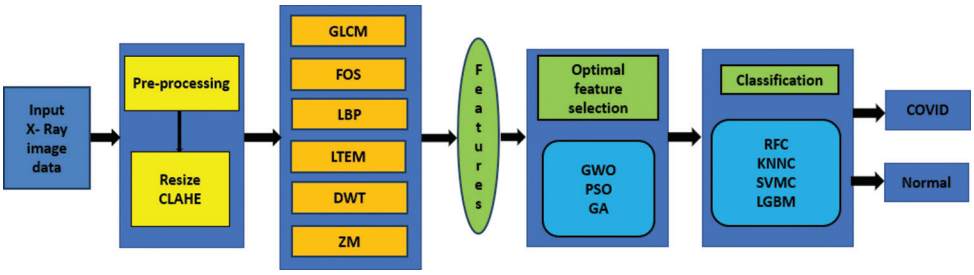

Figure 4: Block diagram representing the working of present study

a total of 98 features, the PSO reduced the features in the range of 47 to 52 from a total of 98 features and the GA reduced the features in the range of 38 to 53 from total 98 features as given in Table 3. On an average, 83 (84.69%) features out of the total features were selected by GWO, 49 (50.00%) features out of the total features were selected by PSO, and on an average, 44 (45.66%) features of the total features were selected by GA in the optimal feature selection.

Classification algorithm comparison with different feature selection algorithms

Table 4 shows the comparative performance of 3 feature selection algorithms and 4 classification algorithms for the classification of 400 CXR images. For the RF classifier, the feature selection algorithms PSO and GA outperform the GWO. For KNN, SVM, and LGB classifiers, the feature selection algorithms PSO outperform the GWO and GA. The python source code is available on https://github.com/MohiniManav/Feature-optimization.

Figure 5 demonstrates the classification performance of classification models on test data with different feature selection algorithms.

In order to demonstrate the performance of the proposed framework over the existing state-of-the-art in terms of various evaluation metrics, a comprehensive performance comparison with other similar works in the literature has been performed. Khuzani *et al*. performed a study with spatial (Texture, GLDM, GLCM), frequency (Wavelet and FFT), and statistical measurements. In their study features were reduced using PCA and classified using an Artificial Neural Network (ANN). In their study, the highest overall accuracy and F1-score values were 95% and 94.3%, respectively.[19]

Singh *et al*. extracted the features from CXR images and relevant features were selected using the Hybrid Social Group Optimization algorithm. With these selected features, a classification accuracy of 99.65% was achieved using the SVM classifier in their study.[58]

Mostafiz *et al*. proposed an approach to detect COVID-19 with good accuracy from the CXR image using the hybridization

Table 3: Feature selection for various classification models using feature selection algorithms

Models used for classification	Methods used for feature reduction	The initial number of features	Selected optimized number of features
RF	GWO	98	83
	PSO	98	48
	GA	98	38
KNN	GWO	98	85
	PSO	98	47
	GA	98	38
SVM	GWO	98	81
	PSO	98	50
	GA	98	53
LGB	GWO	98	83
	PSO	98	52
	GA	98	50

RF: Random Forest, KNN: K-Nearest Neighbour, SVM: Support Vector Machine, GWO: Grey Wolf Optimization, PSO: Particle Swarm Optimization, GA: Genetic Algorithm, LGB: Light Gradient Boosting

Table 4: Performance of different feature selection algorithms with different classifiers

Models used for classification	Methods used for feature reduction	Accuracy	Precision	Recall	F1-score
RF	GWO	88	85	85	85
	PSO	96	94	97	96
	GA	96	94	97	96
KNN	GWO	84	77	88	82
	PSO	93	89	94	91
	GA	90	86	91	89
SVM	GWO	96	92	100	96
	PSO	100	100	100	100
	GA	99	97	100	99
LGB	GWO	93	91	91	91
	PSO	95	92	97	94
	GA	93	89	94	91

RF: Random Forest, KNN: K-Nearest Neighbour, SVM: Support Vector Machine, GWO: Grey Wolf Optimization, PSO: Particle Swarm Optimization, GA: Genetic Algorithm, LGB: Light Gradient Boosting

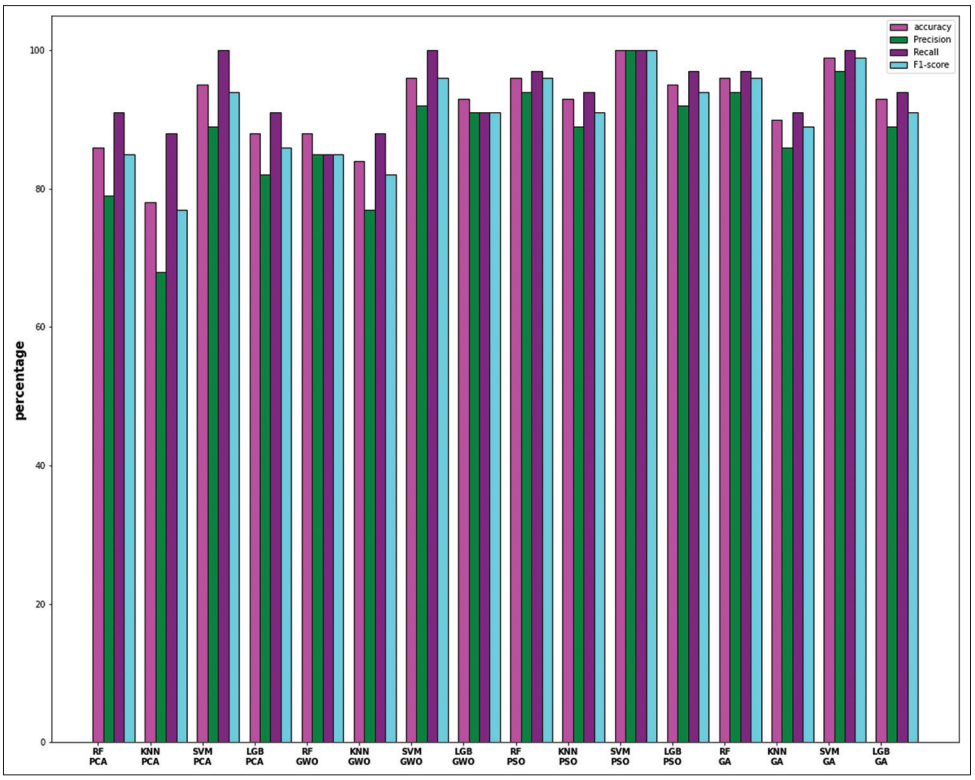

Figure 5: Bar graph representing the classification performance of classification models on test data with different feature selection algorithms

of deep CNN and DWT features. They extracted optimum features from these hybridized features through minimum redundancy and maximum relevance along with RFE and achieved an overall accuracy of more than 98.5%.[59]

Nour *et al*. conducted a 3-class classification study for the classification of COVID-19-normal-viral Pneumonia. They obtained the features with CNN models and optimized the parameters of the models using the Bayesian optimization algorithm. They used SVM and KNN for classification. In their study, the most efficient results were ensured by the SVM classifier with an accuracy of 98.97%, a sensitivity of 89.39%, a specificity of 99.75%, and an F-score of 96.72%.[60] Sahlol *et al*. proposed an improved hybrid classification approach for COVID-19 images by combining the strengths of CNNs to extract features and a swarm-based feature selection algorithm (Marine Predators Algorithm) to select the most relevant features from 2 datasets. They achieved 98.7%, 98.2% and 99.6%, 99% of classification accuracy and F-Score for dataset 1 and dataset 2 respectively.[61] Dias Júnior *et al*. extracted the features from CXR using a deep features-based approach implemented through the networks VGG19, Inception-v3, and ResNet50 and the classified the CXR images into COVID-19 and Non-COVID-19 groups, using eXtreme Gradient Boosting (XGBoost) optimized by PSO. They achieved an accuracy of 98.71%, a precision of 98.89%, a recall of 99.63%, and an F1-score of 99.25%.[62] Mohammed *et al*. used Binary PSO to optimize the LBP features extracted from the CXR images. For the classification of CXR into COVID and Non-COVID category, they have used SVM and KNN classifiers. Their experimental results showed an average accuracy of 94.6%, sensitivity of 96.2%, and specificity equal to 93% with the SVM classifier.[63]

In the present study, for all the feature selection methods, SVM classifier gave the most accurate and precise result as compared to other classification models. Also, in feature selection methods, PSO gave the best result as compared to other methods for feature selection. Using the combination of SVM classifier with the features selected with PSO algorithm, we observed that the accuracy, precision, recall, and F1-score were 100%.

Conclusion

The present study used the optimal number of features extracted from the CXR images for classification. Based on these features, CXR images were classified into two categories, i.e., COVID-19 and normal images. For optimal feature selection, three nature-inspired algorithms GWO, PSO, and GA were used. For classification, RF, KNN, SVM, and LGBM classifiers were used. With optimal features obtained with PSO, the SVM classifier achieved the highest accuracy, precision, recall, and F1 score as compared to others. The result of the study indicates that with optimal features and the best choice of the classifier algorithm, the most accurate computer-aided diagnosis of CXR can be achieved.

Since the present study uses publicly available data so the diversity of the databases can simulate the clinical routine, and makes the classification method amenable to comparison. It

does not require large quantities of hardware resources, which is one of the biggest problems faced in deep learning-based methods. As it is extremely difficult to manually select a feature from the dataset on which the algorithm can perform better, so understanding the dataset and selecting meaningful features from it prior to feeding data into the machine learning algorithm is necessary. The feature selection impacts the final accuracy of the machine learning models. Contrarily, a deep learning algorithm with a small dataset has a high chance of over-fitting. It also requires high computational power, a large feature set, and resources compared to machine learning models.

However, our method did not propose a new machine-learning model. We used the existing models and demonstrated their effectiveness for feature classification. It is believed that developing a new model or modifying the existing classification architectures will further improve results and can be extended as future work.

The minimal consumption of data storage, processing time, and hardware is an additional advantage of the optimal feature selection. Thus, the proposed computer-aided diagnosis approach with optimal features may be used as a complementary tool to assist the radiologist in making a more accurate and early diagnosis of disease using CXR images.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Tracking SARS-CoV-2 Variants. Available from: https://www.who.int/activities/tracking-SARS-CoV-2-variants. [Last accessed on 2022 Nov 14].
2. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): A review. JAMA 2020;324:782-93.
3. Shortage of Personal Protective Equipment Endangering Health Workers Worldwide. Available from: https://www.who.int/news/item/03-03-2020-shortage-of-personal-protective-equipment-endangering-health-workers-worldwide. [Last accessed on 2022 Nov 14].
4. Cascella M, Rajnik M, Aleem A, Dulebohn SC, Di Napoli R. Features, evaluation, and treatment of coronavirus (COVID-19). In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2022. Available From: http://www.ncbi.nlm.nih.gov/books/NBK554776/. [Last accessed on 2022 Nov 14].
5. Xu M, Wang D, Wang H, Zhang X, Liang T, Dai J, *et al.* COVID-19 diagnostic testing: Technology perspective. Clin Transl Med 2020;10:e158.
6. Kumar KS, Mufti SS, Sarathy V, Hazarika D, Naik R. An update on advances in COVID-19 laboratory diagnosis and testing guidelines in India. Front Public Health 2021;9:568603.
7. Chandra TB, Verma K. Pneumonia detection on chest X-ray using machine learning paradigm. In: Chaudhuri BB, Nakagawa M, Khanna P, Kumar S, editors. Proceedings of 3rd International Conference on Computer Vision and Image Processing Singapore (Advances in Intelligent Systems and Computing). Vol. 1022. Singapore: Springer 2020. p. 21-33. Available from: http://link.springer.com/10.1007/978-981-32-9088-4_3. [Last accessed on 2022 Nov 14].
8. Hosseiny M, Kooraki S, Gholamrezanezhad A, Reddy S, Myers L. Radiology perspective of coronavirus disease 2019 (COVID-19): Lessons from severe acute respiratory syndrome and middle east respiratory syndrome. AJR Am J Roentgenol 2020;214:1078-82.
9. Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH. Essentials for radiologists on COVID-19: An update-radiology scientific expert panel. Radiology 2020;296:E113-4.
10. Bohr A, Memarzadeh K. Chapter 2 - The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare [Internet]. Academic Press; 2020. p. 25–60. Available from: https://www.sciencedirect.com/science/article/pii/B9780128184387000022. [Last accessed on 2023 May 20].
11. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, *et al.* AI applications to medical images: From machine learning to deep learning. Phys Med 2021;83:9-24.
12. Bhargava A, Bansal A. Novel coronavirus (COVID-19) diagnosis using computer vision and artificial intelligence techniques: A review. Multimed Tools Appl 2021;80:19931-46.
13. Majumder D. Development of a fast Fourier transform-based analytical method for COVID-19 diagnosis from chest X-ray images using GNU octave. J Med Phys 2022;47:279-86.
14. Nanni L, Ghidoni S, Brahnam S. Handcrafted versus. Non-handcrafted features for computer vision classification. Pattern Recognit 2017;71:158-72.
15. Camilleri D, Prescott T. Analysing the limitations of deep learning for developmental robotics. In: Mangan M, Cutkosky M, Mura A, Verschure PF, Prescott T, Lepora N, editors. Biomimetic and Biohybrid Systems (Lecture Notes in Computer Science). Vol. 10384. Cham: Springer International Publishing; 2017. p. 86-94. Available from: http://link.springer.com/10.1007/978-3-319-63537-8_8. [Last accessed on 2022 Nov 14].
16. Otokiti AU. Digital health and healthcare quality: A primer on the evolving 4th industrial revolution. In: Stawicki SP, Firstenberg MS, editors. Contemporary Topics in Patient Safety - Volume 1. London: IntechOpen; 2022. p. 1-20.
17. Barhate N, Bhave S, Bhise R, Sutar RG, Karia DC. Reducing Overfitting in Diabetic Retinopathy Detection using Transfer Learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA). Greater Noida, India: IEEE; 2020. p. 298-301. Available from: https://ieeexplore.ieee.org/document/9250772/. [Last accessed on 2022 Nov 14].
18. Barhate N, Bhave S, Bhise R, Sutar RG, Karia DC. Reducing Overfitting in Diabetic Retinopathy Detection using Transfer Learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) [Internet]. Greater Noida, India: IEEE; 2020. p. 298-301. Available from: https://ieeexplore.ieee.org/document/9250772/. [Last accessed on 2022 Nov 14].
19. Zargari Khuzani A, Heidari M, Shariati SA. COVID-classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images. Sci Rep 2021;11:9887.
20. Pereira RM, Bertolini D, Teixeira LO, Silla CN Jr., Costa YM. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput Methods Programs Biomed 2020;194:105532.
21. Malik H, Fatema N, Iqbal A. Chapter 1 - Advances in Machine Learning and Data Analytics. In: Malik H, Fatema N, Iqbal A, editors. Intelligent Data-Analytics for Condition Monitoring [Internet]. Academic Press; 2021. p. 3-29. Available from: https://www.sciencedirect.com/science/article/pii/B9780323855105000016. [Last accessed on 2023 May 17].
22. Mangal A, Holm EA. A Comparative study of feature selection methods for stress hotspot classification in materials. Integr Mater Manuf Innov 2018;7:87-95.
23. Sutter JM, Kalivas JH. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. Microchem J 1993;47:60-6.
24. Hamada M, Tanimu JJ, Hassan M, Kakudi HA, Robert P. Evaluation of recursive feature elimination and LASSO regularization-based optimized feature selection approaches for cervical cancer prediction. In: 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC) Singapore. Singapore: IEEE; 2021. p. 333-9. Available from: https://ieeexplore.ieee.org/document/9691991/. [Last accessed on 2022 Nov 14].

25. Song F, Mei D, Li H. Feature selection based on linear discriminant analysis. In: 2010 International Conference on Intelligent System Design and Engineering Application. Changsha, Hunan, China: IEEE; 2010. p. 746-9. Available from: http://ieeexplore.ieee.org/document/5743287/. [Last accessed on 2022 Nov 14].
26. Chandra TB, Verma K, Singh BK, Jain D, Netam SS. Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble. Expert Syst Appl 2021;165:113909.
27. Öztürk Ş, Özkaya U, Barstuğan M. Classification of coronavirus (COVID-19) from X-ray and CT images using shrunken features. Int J Imaging Syst Technol 2021;31:5-15.
28. Bhargava A, Bansal A, Goyal V. Machine learning-based automatic detection of novel coronavirus (COVID-19) disease. Multimed Tools Appl 2022;81:13731-50.
29. Kumar R, Arora R, Bansal V, Sahayasheela VJ, Buckchash H, Imran J, *et al.* Classification of COVID-19 from chest x-ray images using deep features and correlation coefficient. Multimed Tools Appl 2022;81:27631-55.
30. Sethy PK, Behera SK, Ratha PK, Biswas P. Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. Int J Math Eng Manage Sci 2020;5:643-51.
31. Tahir AM, Chowdhury ME, Khandakar A, Rahman T, Qiblawey Y, Khurshid U, *et al.* COVID-19 infection localization and severity grading from chest X-ray images. Comput Biol Med 2021;139:105002.
32. COVID-QU-Ex Dataset. Available from: https://www.kaggle.com/datasets/cf77495622971312010dd5934ee91f07ccbcfdea8e2f7778977ea8485c1914df. [Last accessed on 2022 Nov 15].
33. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Abul Kashem SB, *et al.* Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput Biol Med 2021;132:104319.
34. Degerli A, Ahishali M, Yamac M, Kiranyaz S, Chowdhury ME, Hameed K, *et al.* COVID-19 infection map generation and detection from chest X-ray images. Health Inf Sci Syst 2021;9:15.
35. Chowdhury ME, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, *et al.* Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 2020;8:132665-76.
36. Shaziya H, Shyamala K, Zaheer R. Comprehensive review of automatic lung segmentation techniques on pulmonary CT images. In: 2019 Third International Conference on Inventive Systems and Control (ICISC). Coimbatore, India: IEEE; 2019. p. 540-5. Available from: https://ieeexplore.ieee.org/document/9036429/. [Last accessed on 2022 Nov 15].
37. Pizer SM, Johnston RE, Ericksen JP, Yankaskas BC, Muller KE. Contrast-limited adaptive histogram equalization: speed and effectiveness. In: (1990) Proceedings of the First Conference on Visualization in Biomedical Computing. Atlanta, GA, USA: IEEE Computer Society Press; 1990. p. 337-45. Available from: http://ieeexplore.ieee.org/document/109340/. [Last accessed on 2022 Nov 15].
38. Aggarwal NK, Agrawal R. First and second order statistics features for classification of magnetic resonance brain images. JSIP 2012;3:146-53.
39. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst, Man, Cybern. 1973;SMC-3:610-21.
40. Galloway MM. Texture analysis using gray level run lengths. Comput Graph Image Process 1975;4:172-9.
41. Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of 12th International Conference on Pattern Recognition. Jerusalem, Israel: IEEE Computer Society Press; 1994. p. 582-5. Available from: http://ieeexplore.ieee.org/document/576366/. [Last accessed on 2022 Nov 15].
42. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 1996;29:51-9.
43. Laws KI. Rapid Texture Identification. In: Image Processing for Missile Guidance. San Diego, California.Vol.288. Bellingham, WA: Society of Photo-optical Instrumentation Engineers; 1980. p. 376-81.
44. Ghazali KH, Mansor MF, Mustafa Mohd M, Hussain A. Feature extraction technique using discrete wavelet transform for image classification. In: 2007 5th Student Conference on Research and Development. Selangor, Malaysia: IEEE; 2007. p. 1-4. Available from: http://ieeexplore.ieee.org/document/4451366/. [Last accessed on 2022 Nov 15].
45. Teague MR. Image analysis via the general theory of moments*. J Opt Soc Am 1980;70:920.
46. Giakoumoglou NG. Pyfeats: Image Feature Extraction Inside Region-of-Interest. Available from: https://github.com/giakou4/pyfeats. [Last accessed on 2022 Nov 15].
47. Peshawa MJ, Faraj RH. Data normalization and standardization: A technical report.Mach Learn Tech Rep. 2014;1:1-6.
48. Al-Thanoon NA, Qasim OS, Algamal ZY. Improving nature-inspired algorithms for feature selection. J Ambient Intell Hum Comput 2022;13:3025-35.
49. Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. Adv Eng Softw 2014;69:46-61.
50. Tran B, Xue B, Zhang M. Overview of particle swarm optimisation for feature selection in classification. In: Dick G, Browne WN, Whigham P, Zhang M, Bui LT, Ishibuchi H, *et al*., editors. Simulated Evolution and Learning (Lecture Notes in Computer Science. Vol. 8886. Cham: Springer International Publishing; 2014. p. 605-17. Available from: http://link.springer.com/10.1007/978-3-319-13563-2_51. [Last accessed on 2022 Nov 16].
51. Hussein F, Kharma N, Ward R. Genetic algorithms for feature selection and weighting, a review and study. In: Proceedings of Sixth International Conference on Document Analysis and Recognition. Seattle, WA, USA: IEEE Computer Society; 2001. p. 1240-4. Available from: http://ieeexplore.ieee.org/document/953980/. [Last accessed on 2022 Nov 16].
52. Development GitHub. Available from: https://github.com/jaswinder9051998/zoofs. [Last accessed on 2022 Nov 17].
53. Breiman L. Random forests. Mach Learn 2001;45:5-32.
54. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13:21-7.
55. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273-97.
56. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, *et al*., editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017.
57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825-30.
58. Singh AK, Kumar A, Mahmud M, Kaiser MS, Kishore A. COVID-19 Infection Detection from Chest X-Ray Images Using Hybrid Social Group Optimization and Support Vector Classifier. Cognit Comput. 2021;1-13.
59. Mostafiz R, Uddin MS, Alam NA, Reza M, Rahman MM. COVID-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features. J King Saud Univ Comput Inf Sci 2022;34 (6, Part B):3226-35.
60. Nour M, Cömert Z, Polat K. A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization. Appl Soft Comput. 2020;97:106580.
61. Sahlol AT, Yousri D, Ewees AA, Al-Qaness MA, Damasevicius R, Elaziz MA. COVID-19 image classification using deep features and fractional-order marine predators algorithm. Sci Rep 2020;10:15364.
62. Dias Júnior DA, da Cruz LB, Bandeira Diniz JO, França da Silva GL, Junior GB, Silva AC, *et al.* Automatic method for classifying COVID-19 patients based on chest X-ray images, using deep features and PSO-optimized XGBoost. Expert Syst Appl 2021;183:115452.
63. Mohammed BN, Al-Mukhtar FH, Yousif RZ, Almashhadani YS. Automatic classification of COVID-19 chest X-ray images using local binary pattern and binary particle swarm optimization for feature selection. Cihan Univ Erbil Sci J 2021;5:46-51.

Supplementary Table 1: Extracted features from X-rays

Feature extracted from	Features extracted (features name given here as mentioned in zoofs python library)
FOS features	"FOS_Mean," "FOS_Variance," "FOS_Median," "FOS_Mode," "FOS_Skewness," "FOS_Kurtosis," "FOS_Energy," "FOS_Entropy," "FOS_MinimalGrayLevel," "FOS_MaximalGrayLevel," "FOS_CoefficientOfVariation," "FOS_10Percentile," "FOS_25Perce-ntile," "FOS_75Percentile," "FOS_90Percentile," "FOS_HistogramWidth"
GLCM	"GLCM_ASM," "GLCM_Contrast," "GLCM_Correlation," "GLCM_SumOfSquaresVariance," "GLCM_InverseDifferenceMoment," "GLCM_SumAverage," "GLCM_SumVariance," "GLCM_SumEntropy," "GLCM_Entropy," "GLCM_DifferenceVariance," "GLCM_DifferenceEntropy," "GLCM_Information1," "GLCM_Information2," "GLCM_MaximalCorrelationCoefficient"
LTE	"LTE_LL_7," "LTE_EE_7," "LTE_SS_7," "LTE_LE_7," "LTE_ES_7," "LTE_LS_7"
GLRLM	"GLRLM_ShortRunEmphasis," "GLRLM_LongRunEmphasis," "GLRLM_GrayLevelNo-Uniformity," "GLRLM_RunLengthNonUniformity", "GLRLM_RunPercentage", "GLRLM_LowGrayLevelRunEmphasis", "GLRLM_HighGrayLevelRunEmphasis," "GLRLM_Short owGrayLevelEmphasis," "GLRLM_ShortRunHighGrayLevelEmphasis," "GLRLM_LongRunLowGrayLevelEmphasis," "GLRLM_LongRunHighGrayLevelEmphasis"
LBP	"LBP_R_1_P_8_energy," "LBP_R_1_P_8_entropy," "LBP_R_2_P_16_energy," "LBP_R_2_P_16_entropy," "LBP_R_3_P_24_energy," "LBP_R_3_P_24_entropy"
DWT	"DWT_bior3.3_level_1_da_mean," "DWT_bior3.3_level_1_da_std," "DWT_bior3.3_level_1_dd_mean," "DWT_bior3.3_level_1_dd_std," "DWT_bior3.3_level_1_ad_mean," "DWT_bior3.3_level_1_ad_std," "DWT_bior3.3_level_2_da_mean," "DWT_bior3.3_level_2_da_std," "DWT_bior3.3_level_2_dd_mean," "DWT_bior3.3_level_2_dd_std," "DWT_bior3.3_level_2_ad_mean," "DWT_bior3.3_level_2_ad_std," "DWT_bior3.3_level_3_da_mean," "DWT_bior3.3_level_3_da_std," "DWT_bior3.3_level_3_dd_mean," "DWT_bior3.3_level_3_dd_std," "DWT_bior3.3_level_3_ad_mean," "DWT_bior3.3_level_3_ad_std"
ZM	"zenikes_0," "zenikes_1," "zenikes_2," "zenikes_3," "zenikes_4," "zenikes_5," "zenikes_6," "zenikes_7," "zenikes_8," "zenikes_9," "zenikes_10," "zenikes_11," "zenikes_12," "zenikes_13," "zenikes_14," "zenikes_15," "zenikes_16," "zenikes_17," "zenikes_18," "zenikes_19," "zenikes_20," "zenikes_21," "zenikes_22," "zenikes_23," "zenikes_24"

FOS: First-order statistics, GLCM: Gray-level co-occurrence matrix, GLRLM: Gray-level run length matrix, LBP: Local binary pattern, DWT: Discrete wavelet transform, ZM: Zernike's moments, LTE: Law's Texture Energy