ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
February 2020, Vol 13(07), 805 – 816
DOI: 10.17485/ijst/2020/v13i07/149846,

Extraction with Map-Reduce Framework and Correlation-based Feature Selection in Lung Cancer Towards Big Data

R. Sujitha* and V. Seenivasagam

Department of Information Technology, National Engineering College (Autonomous), Kovilpatti, Tamilnadu, India

Article Type: Article

Article Citation: Sujitha R, Seenivasagam V. Extraction with map-reduce framework and correlation-based feature selection in lung cancer towards big data. *Indian Journal of Science and Technology*. 2020; 13(07),805-816.DOI:10.17485/ijst/2020/v013i07/149846

Received date: January 8, 2020

Accepted date: January 29, 2020

***Author for correspondence:** R. Sujitha sujitha.r24@gmail.com Department of Information Technology, National Engineering College (Autonomous), Kovilpatti, Tamilnadu, India

Abstract

Background/objectives: To extract nucleus and cytoplasm that intend to optimize features in high-dimensional images such as all types of raw sputum cells. To calculate following features efficiently: Area, Perimeter, Intensity, NC Ratio, and Circularity. **Methods/Statistical analysis**: To take results in proposed stride, we introduced map-reduce framework for separating similar cells from sputum cell images that have been collected from Microscope lab images with intended magnification and staining. To avoid model learn from irrelevant features, feature selection methods with correlation-based feature selection contributes appropriate features that are then fed for classification. Features here converted to vectors for the estimation of symmetric uncertainty, correlation-based approach. **Findings**: Performance evaluation metrics checks into the contribution to measure it's out coming performance. Even though lot of works relied on feature extraction, our work combines feature extraction with map-reduce framework which improves accuracy for classification. Our proposed method makes extraction of nucleus and cytoplasm easier than other methods. Optimized performance assured in proposed feature selection. **Novelty/applications**: Eventual accuracy for every feature in proposed stride improves than other existing works. In addition, ROC curves proves higher true positive rate even in increased datasets. Another significant innovation in our work is map-reduce framework applies in images to sort cells with respect to staining.

Keywords: Health Care, Correlation, Classification, Big Data, Map-reduce, Sputum.

1. Introduction

Lung cancer confronts with lots of people rapidly. The contentious characteristic of lung cancer depends on its treatment and its outcomes. Because many types of lung cancer

grow quickly and spread rapidly and the lungs are vital organs, early detection and prompt treatment—usually surgery to remove the tumor—is critical. Medical diagnosis prompts several approaches to detect and cure lung cancer. Computed tomography images and sputum cell images engaged with lung cancer prediction and classification. The quality of images perhaps becomes fascinating features to predict lung cancer, which achieves through some image processing techniques. The mode of classifying lung cancer proceeds with the extraction of features such as area, perimeter, eccentricity followed by feature selection methods. Even though the features are categorical nature, some features extract larger values irrelevant to expected outcomes. Hence optimal feature selection techniques proposed. Meanwhile, inconsistency removes by underlying map-reduce framework in various types of sputum images such as eosinophilia, bronchial mucus, squamous carcinoma cells and then fed to feature extraction using MATLAB. Furthermore, feature selection and classification work in the ML-PYSPARK environment for parallel processing of a large number of datasets.

In Ref. [1], Lin proposes an approach for optimal analysis of feature importance and as effective classification, a PQD recognition method based on image enhancement techniques. Furthermore, the Gini Index used to evaluate sequence forward search and for optimal feature subset selection. Meanwhile, Disturbance features extracted and eliminated from binary image and images are reconstructed for classification purposes. In Ref. [2], Wang evaluates an approach for structural discrimination of networks and also to estimate discriminate features in brain disease classification, graph kernel-based structured feature selection (gk-SFS). Meanwhile, to improve performance, l_1-norm based sparsity regularizer deployed in [2]. Performance achieves by comparing the accuracy of proposed one with other existing approaches, as well as considers eliminating noisy or redundant information in [2]. In Ref. [3], Chen provides an approach based on incurring security over image manipulation detection attacks. They address random feature selection, which incorporates negligible loss of performance. Specifically, they focus on two issues as adaptive histogram equalization and median filtering which consolidates effective manipulation techniques. In Ref. [4], Nie proposes a method for selecting non-redundant and representative features, an auto-weighted feature selection framework via global redundancy minimization (AGRM) which is non-parametric and is weighing automatically. Hence, they address both supervised and non-supervised feature selection. Meanwhile, our proposed map-reduce framework illustrates the author of Ref. [5]. Furthermore, they provide feature scores for redundant features and compares performance over other existing approaches. In Ref. [6], Singh proposed a novel filter-based approach for feature selection that sorts out the features based on a score. The proposed framework layout abruptly improves the results, even with high-dimensional datasets. Moreover, they tried to improve performance over other classification accuracy and precision. In Ref. [7], Maulik proposed a prediction scheme that combines a fuzzy preference-based rough set (FPRS) method for feature (gene) selection with semi-supervised SVMs. Even more, they have shown effectiveness by comparing with the signal-to-noise ratio (SNR) and consistency based feature selection (CBFS) methods. In Ref. [8], Shen depicts methodology that combines fused lasso and elastic net as regularization for linear support vector machine (SVM), also called feature selection SVM (OFSSVM),

which uses huberized hinge loss as the loss function. However, the author improves performance by adding both binary as well as Multi-class classification. In Ref. [9], Zhang predicts the prognosis and survival time of different subtypes of GBM by introducing combined gene testing with clinical treatment and extracts dataset from Cancer Genome Atlas (TCGA) database, which further improves the efficiency of standards. In this way, they depict the minimum redundancy feature selection method (mRMR) and the Multiple Kernel Machine (MKL) learning method for effective prediction and feature selection. In Ref. [10], Taşkın proposed the classification task which addresses dimensionality reduction by improving the correlation between the spectral features and the noise present in spectral bands. In addition, performance improved by proving the stability of the feature selection method and computational time of classification as well as accuracy. Specifically, they have used hyper spectral datasets as images. In Ref. [11], Archibald proposed a band selection method that co-occurs to assist with classification accuracy. In addition, an embedded-feature-selection (EFS) algorithm that is tailored to operate with support vector machines (SVMs) introduce to perform band selection and classification in parallel to reduce the computational time which further converges its performance. In Ref. [12], Chong depicts Robustness-Driven Feature Selection (RDFS) algorithm that dramatically increases robustness in CT images, considering various factors. In addition, two SVM-based approaches, one with RDFS and another without RDFS to coherently compare the robustness of the proposed algorithm. Moreover, the comparison performed in the multi-reconstruction dataset, using Cohen's kappa classification factor. In Ref. [13], Xiabi Liu et al. introduced a novel approach of fisher criterion and genetic optimization (FIG) which selects subsets of various features considering some factors including bag-of-visual-words based on the histogram of oriented gradients, the wavelet transform-based features, the local binary pattern, and the CT value histogram. In Ref. [14], Bolourchi proposed score-based approach, entropy score selecting top k methods in Synthetic Aperture radar images for dimensionality reduction and to achieve feasibility in feature selection. High-dimensional datasets used by combining various sets into vectors and improves performance of accuracy on image classification. In Ref. [15], Fauvel proposed a framework for hyper spectral images, Gaussian mixture model classifier (GMM).They classify images and improves performance over the K-folds cross-validation approach and prove the significance of the proposed classification. Besides [16–18], deals with sputum cells for nucleus and cytoplasm extraction and with big data workflow. Furthermore [19], process feature selection for the purpose of better classification. Some more concepts layout machine learning techniques for feature extraction as in Ref. [20]. In Ref. [21], CT images collected and stages of lung cancer determined using the concept SVM in addition with image contrast enhancement and optimal feature extraction techniques.

2. Materials and Methods

2.1. Our Contribution

Map-reduce framework has been developed to put off ill-suited images, by using mapper phase, so as to sort similar type of cells and fed to reducer phase as in Figure 1.

Retrieved images endure feature extraction techniques as in sub-sections.

Moreover, correlation-based symmetric uncertainty technique prospers feature selection to expedite classification.

2.1.1. Feature Extraction

Images have been collected from several government hospitals in Tamil Nadu at magnification 40× with PAP staining and H&E staining. Raw sputum images with various cells such as eosinophilia, bronchial mucous cells, and squamous carcinoma cells are having nucleus and cytoplasm in different nature. Hence, they are processed using the map-reduce framework as in Figure 1 to splits cells with the same nature and then encounters with K-means clustering followed by some morphological operations such as erosion and dilation to individualize the combined nucleus so that, other parameters such as area, perimeter computes accurately.

RGB images are first converted into gray level images to remove the noise. Then the images have been actuated into noise removal. The median filter removes the noise and filters the images with gray images. The gray images are then reconstructed to RGB images using MATLAB code. The RGB images are then fed to L*a*b color space for further processing,

- **L*: Lightness**
- **a*: Red/Green Value**
- **b*: Blue/Yellow Value**

The LAB color model is a three-axis color system. The first axis, the L-channel or Lightness, goes up and down the 3D color model and it consists of white to black – and all of your gray colors will be exactly right down the center. The A-axis goes from cyan color across to magenta/red color and the B axis goes from blue to yellow. Also derived as device-independent; the colors in L*a*B color space have fluctuated to K-means clustering which extracts nucleus from sputum cells. K-means hold images to cluster and extract

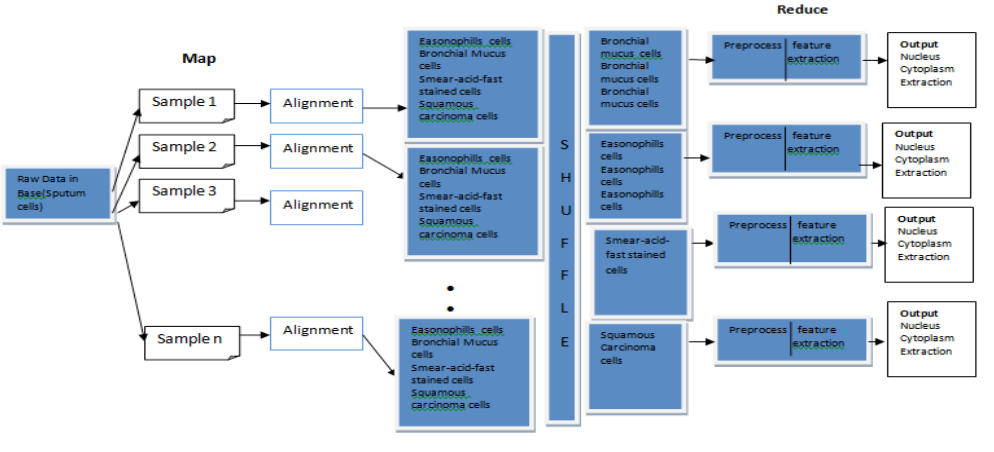

FIGURE 1. Map-reduce framework for images classification.

the targeted features. Iteration for the algorithm used is 3 times. Furthermore, the region growing algorithm as similar to the method proposed by Ref. [16] classified as a pixel-based image segmentation method to select necessary points or region has been applied as in Figure 2(b). After processing with region growing, images are further actuated to K-means clustering which clusters the images and iteratively processes the images. Besides, a morphological operation such as erosion and dilation intrudes to separate connected nuclei.

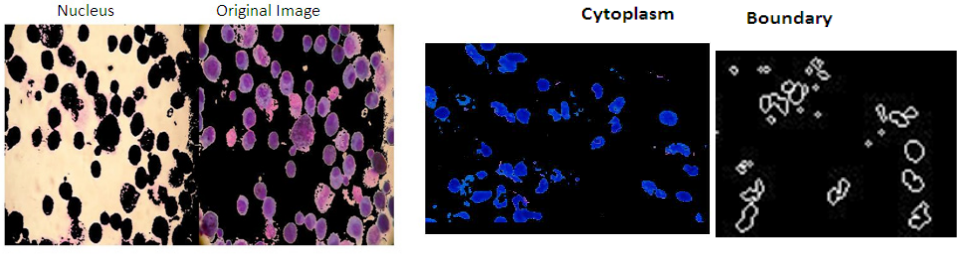

FIGURE 2. a. Nucleus extraction. b. Cytoplasm.

2.1.1.1. Parameters and Features

The first feature is the NC ratio, which is computed by dividing the actual number of pixels over the nucleus region (nuclei area) as in Figure 3 and the cytoplasmic region as in Figure 2.

- *Area*

It is a scalar value which derives an actual number of pixels in nuclei and cytoplasm extraction.

Area calculated as

$$A = \sum_{i}\sum_{j}(A_{i,j}, \text{nucleic area} = i, \text{cytoplasm area} = j) \tag{1}$$

where i, j depicts nucleic area and the cytoplasmic area which contains the number of pixels in both region.

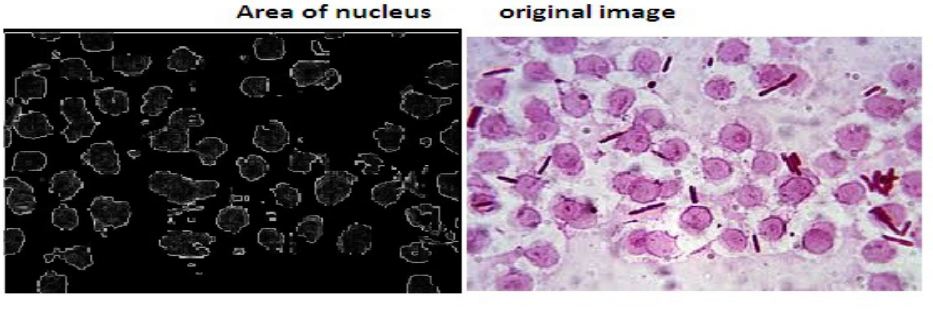

FIGURE 3. Nucleus area.

- *NC Ratio*
  NC ratio evaluated using area of nucleus and cytoplasm. Formula layout as in (1)

$$\text{NC ratio} = \frac{\text{Area of nucleus}}{\text{Area of cytoplasm}} * 100 \tag{2}$$

- *Intensity*
  Intensity disputes collection of color pixels of extracted nucleus and cytoplasm.
- *Mean*

$$\text{Mean}\left(\text{cytoplasm}\right) = \frac{\sum_{i=1}^{N} \text{Intensity}\left(i\right)}{\text{Area}\left(\text{cytoplasm}\right)} \tag{3}$$

$$\text{Mean}\left(\text{nucleus}\right) = \frac{\sum_{i=1}^{N} \text{Intensity(i)}}{\text{Area(nucleus)}} \tag{4}$$

Perimeter

$$P_{i,j}\text{edge}\left(\text{nucelei}\right) = i, \text{edge}\left(\text{cytoplasm}\right) = j \tag{5}$$

where edge (nuclei) and edge (cytoplasm) are the vector co-ordinates of i and j, respectively.

- *Circularity*

$$\text{circularity} = \frac{4\pi * \text{nucleiarea}}{\text{Perimeter}\left(\text{nuclei}\right)^2} \tag{6}$$

2.2. Feature Selection

Feature selection is the process of selecting subspaces with appropriate features that are then used for developing a model. To resolve such problems, feature selection plays several approaches.

2.2.1. Correlation-based Feature Selection

It correlates nominal features that act as a measure of the endowment for features. We focus on ignoring features among those sputum cells which might give relevant and redundant features. Several measures behind correlation-based feature selection such as

1) Relief
2) Minimum Description Length (MDL)
3) Symmetrical uncertainty

2.2.1.1. Relief

Relief computes score or value for each feature and obtains the highest score feature by applying rank for each score. In addition, the score obtained is used as feature weights which are indulged in the model.

Probability of different feature weights calculated as

Relief$_W$ = *P(value of w| neighbor instance of an inter-class) – P(value of w| neighbor instance of intraclass)* (7)

This is then resolved by eliminating conditional attributes and becomes a value of w|intra class or inters class instead of using neighbor instances as in (8).

$$Relief_W = \frac{[\sum_{h\in C} p(h)(1-p(h))] - \sum_{w\in W}(\frac{p(w)^2}{\sum_{w\in W} p(w)^2} \sum_{h\in C} p(h|w)(1-p(h|w)))^* \sum_{w\in W} p(w)^2}{(1-\sum_{h\in C} p(h)^2)\sum_{h\in C} p(h)^2} \quad (8)$$

Whenever Relief used, it measures twice since each feature treated as a class. h is a subset of class C, w is the weight of each feature. Here, the weight of each feature calculated twice as said above.

2.2.1.2. Minimum Description Length (MDL)

A model though reduces description length of data, complex and high cost and its values are too large to compute. Description length of theory, data are approximated as description length of data given theory summed with description length of theory where theory as T, data as D. The description length of data given theory computed by multiplying deterioration (entropy) of B given A. [1] (thesis)

MDL also measures quality with the below equation. The etiquette equation is as follows [2]. (kon95 in the thesis)

$$Pr\_MDL = \log_2 \binom{N}{N_1 \ldots N_C} + \log_2 \binom{N+C-1}{C-1} \quad (9)$$

$$Po\_MDL = \sum_j \log_2 \binom{N_j}{N_{1j}, \ldots, N_{C_j}} + \sum_j \log_2 \binom{N_j+C-1}{C-1} \quad (10)$$

MDL computed with (9) and (10) as

$$MDL = \frac{Pr\_MDL - Po\_MDL}{N} \quad (11)$$

where N is no. of training attributes in class C, N_j numbers of attributes with a jth value of the given attribute, Pr_MDL is prior_MDL and Po_MDL is post_MDL.

2.2.1.3. Symmetric Uncertainty

Symmetric uncertainty derived by determining first the probability of attributes A and B with values $a \in A$, $b \in B$, respectively. The individual probability values of attributes are partitioned concerning other attributes and if values partitioned becomes lesser than non-partitioned values of other attributes, then both attributes are in relation as in [1].

$$Entropy\ value\ of\ A = -\sum_{a \in A} p(a) \log_2 \left(p(a) \right) \tag{12}$$

Formally,

$$Entropy\ value\ of\ B = -\sum_{b \in B} p(b) \log_2 \left(p(b) \right) \tag{13}$$

If as said, there is a relation between attributes A and B, then the equation becomes

$$Entropy\ of\ Bgiven\ A = -\sum_{a \in A} p(a) \sum_{b \in B} p(b|a) \log_2 \left(p(b|a) \right) \tag{14}$$

The quantity about value decrease in the partitioning of B after observing attribute A also measures information gain [1] as follows

$$Info\text{-}Gain = Entropy\ value\ of\ B + Entropy\ value\ of\ A\text{-}Entropy\ (A, B) \tag{15}$$

This Info-Gain is then applied for computation of Symmetric uncertainty as

$$Suc = 2.0^* \frac{\text{Info} - \text{Gain}}{\text{Entropy value of B} + \text{Entropy value of A}} \tag{16}$$

The values are normalized to the range (0, 1) with (16).

We used Symmetric Uncertainty for feature selection since Relief selector uses ranking which gradually changes for every iteration. Meanwhile, MDL cost consuming and values are too large to compute.

3. Results and Discussion

Sputum color images from the microscope lab collected and processed using the map-reduce framework, which bent over backward to make the outcome certainty. Several extraction techniques deployed to retrieve several features. Feature selection is then applied to give benefits of optimized features. Figure 4 depicts features importance value with other feature selection methods. Other feature selection methods include Chi-square, Recursive Feature Elimination (RFE), Random Forest whose nature selects based on filter and wrapper methods without any feature transformation.

Specifically, we have taken certain features and applied chi-square selection as well as proposed Symmetric uncertainty and results are compared as in Figure 5. Moreover, individual results of other methodologies illustrate as in Figure 5.

Out[65]:

	Chi-2	Feature	LightGBM	Logistics	Pearson	RFE	Random Forest	Total
1	True	Perimeter	True	True	True	True	True	6
2	True	Circularity	True	True	True	True	True	6
3	True	Area	True	True	True	True	True	6
4	True	diameter	True	True	True	True	False	5
5	True	centroid	True	True	True	True	False	5
6	True	Mean	True	True	True	True	False	5
7	True	Intensity	True	True	True	True	False	5

FIGURE 4. Analysis of feature values with existing methods.

```
pValues: [0.6872892787909721,0.6822703303362126]
degreesOfFreedom: [2, 3]
statistics: [0.75,1.5]
```

(a)

```
Num Features: 3
Selected Features: [False False False False  True  True False  True]
Feature Ranking: [5 4 6 2 1 1 3 1]
```

(b)

```
Vec 2: [4 5 6 7 4 5 6 7 4 5 6 7 4 5 6 7 4 5 6 7]
Entropy: 2.0
Mutual information: 2.0000000000000004
Symmetrical uncertainty: 0.6327183637591292
```

(c)

FIGURE 5. Screenshots for optimized feature values with various techniques (a) chi-square (b) ranking (c) symmetric uncertainty (proposed).

Since chi-square retrieves degree of freedom as 2, changes in features selection reflect in results. Symmetric Uncertainty removes irrelevant features by calculating values of features with labels and also by setting threshold as 0.60 and ignores features with values less than the threshold.

Symmetric uncertainty method with threshold put off the model with all its might and proves as worthwhile approach as in Table 1.

Sputum color images collected in various microscope labs with intended magnification at 40 xs. Since staining differs for each type of cells, we used PAP staining and H&E staining for cells. To extract nucleus and cytoplasm accordingly, we first develops map-reduce framework which sorts similar type of cells in reducer phase in inundate manner. From our perspective view, we furnish extraction and optimal feature selections in the sputum color images with map-reduce framework and pyspark, respectively. Experimental results show that the proposed model persists better than other existing models. Even though several features such as NC ratio, mean, circularity, intensity are taken into consideration in other works, we made our contribution, an expedite approach in high dimensional data sets by using map-reduce framework. Table 1 shows that features importance in terms of accuracy. This approach obtained overall accuracy of 91% as in Table 1 which is higher than the values reported in the related works. In the literature, some works extract features using various MATLAB properties. Our work focused on separating the features over

TABLE 1. Accuracy measures with existing works

S. no	Algorithms	Accuracy				
		Features				
		NC ratio	Area	Perimeter	Circularity	Intensity
1.	Graph-based Kernel [12]	82%	90%	92%	84%	85%
2.	Computer-aided diagnosis [18]	84%	93%	90%	85%	87%
3.	Proposed K-means and symmetric*	85%	95%	91%	85%	86%

*depicts proposed

the map-reduce framework in high-dimensional datasets as well as obvious and valuable results for further classification.

Furthermore, Figure 6 shows that our proposed model maintains improved true positive rate, even in an unpleasant situation. Our work compares existing models to prove accuracy as in Table 1.

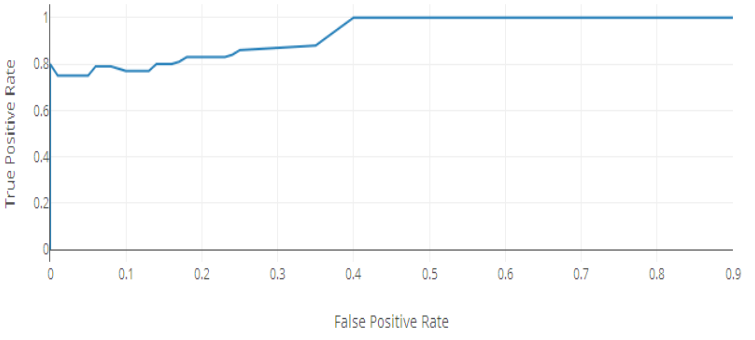

FIGURE 6. ROC curve for optimized feature selection.

4. Conclusion

This work enumerates feature extraction over the map-reduce framework as well as optimal correlated feature selection fair to middling classification better. To this end, sputum color images are given as features for processing over map-reduce as well as correlated symmetric uncertainty feature selection which intends feature importance to improve performance. The filters used are median and wiener filter which removes noise and results from filters are then converted back to RGB images for better extraction of color images. Furthermore, feature subset selection optimized by converting them to dense vectors and endures with symmetric uncertainty methods. Even more, already said approaches, provide optimal feature selection techniques with randomized and preprocessed images. Our proposed techniques are derived with a map-reduce framework and also with raw sputum images including various cells as the source. We enhance the performance over ROC curves and

illustrate uniqueness in demonstrating sputum images in the map-reduce framework. Our future work focuses on a security-based intellectual approach for efficient classification in both binary as in and multi-class classification.

Acknowledgement

The authors sincerely thank anonymous reviewers' constructive comments. The work has been supported by the National Engineering College R&D Program of India and management of National Engineering College.

References

1. Lin L. Power quality disturbance feature selection and pattern recognition based on image enhancement techniques. *IEEE Access*. 2019, 7, 1–16. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8718633
2. Wang M. Graph-kernel based structured feature selection for brain disease classification using functional connectivity networks. *IEEE Access*. 2019, 7, 35001–35011. https://doi.org/10.1109/ACCESS.2019.2903332
3. Chen Z. Secure detection of image manipulation by means of random feature selection. *IEEE Transactions on Information Forensics and Security*. 2019, 14(9), 2454–2469. https://doi.org/10.1109/TIFS.2019.2901826
4. Nie F. A General framework for auto-weighted feature selection via global redundancy minimization. *IEEE Transactions on Image Processing*. 2019, 28(5). https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1109%2FTIFS.2019.2901826
5. Kouanoua AT. An optimal big data workflow for biomedical image analysis. *Informatics in Medicine*. 2018, 68–74. https://doi.org/10.1016/j.imu.2018.05.001
6. Singh B. A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty. *Journal of Data Analysis and Information Processing*. 2014, 2, 95–105. http://dx.doi.org/10.4236/jdaip.2014.24012
7. Maulik U. Fuzzy preference-based feature selection and semi-supervised SVM for cancer classification. *IEEE Transactions on Nano Bioscience*. 2014, 13(2). https://doi.org/10.1109/TNB.2014.2312132
8. Shen Y. Oriented feature selection SVM applied to cancer prediction in precision medicine, special section on big data learning and discovery. *IEEE Access*. 2018. DOI: 10.1109/ACCESS.2018.2868098.
9. Zhang Y. Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016, 13(5). https://doi.org/10.1109/TCBB.2016.2551745
10. Taşkın G. Feature selection based on high dimensional model representation for hyperspectral images. *IEEE Transactions on Image Processing*. 2017, 26(6), 2918–2928. https://doi.org/10.1109/TIP.2017.2687128
11. Archibald R. Feature selection and classification of hyperspectral images with support vector machines. *IEEE Geoscience and Remote Sensing Letters*. 2007, 4(4), 674–677. https://doi.org/10.1109/LGRS.2007.905116

12. Chong DY. Robustness-driven feature selection in classification of fibrotic interstitial lung Disease patterns in computed tomography using 3D texture features. *IEEE Transactions on Medical Imaging*. 2016, 35(1), 144–157. DOI: 10.1109/TMI.2015.2459064.
13. Xiabi Liu, et.al. Recognizing common CT imaging signs of lung diseases through a new feature selection method based on Fisher criterion and genetic optimization. *Journal of Biomedical and Health Informatics*. 2015, 19(2), 144–157. DOI: 10.1109/TMI.2015.2459064.
14. Bolourchi P. Entropy-score-based feature selection for moment-based SAR image classification. *Electronics Letters*. 2018, 54(9), 593–595.https://doi.org/10.1049/el.2017.4419
15. Fauvel M. Fast forward feature selection of hyperspectral images for classification with Gaussian mixture models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2015, 8(6), 2824–2831. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1109%2FJSTARS.2015.2441771
16. Sarrafzadeh O, Dehnavi AM.Nucleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing. *Advanced Biomedical Research*. 2015, 4, 174. DOI: 10.4103/2277-9175.163998.
17. Taher F. Rule based classification of sputum images for early lung cancer detection. *IEEE*. 2016. DOI: 10.1109/ICECS.2015.7440241.
18. Taher F. Computer aided diagnosis system for early lung cancer detection. In: International conference on systems, signals and image processing. 2015. https://doi.org/10.1109/IWSSIP.2015.7313923
19. Sun BY. Combined feature selection and prognosis using support vector progression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 2011, 8(6), 1671–1677. DOI: 10.1109/TCBB.2010.119.
20. Ummadi J, Reddy, Venkata B, Reddy R. LESH -feature extraction and cognitive machine learning techniques for recognition of lung cancer cells. 2019. https://www.researchgate.net/publication/335029136_LESH_-feature_extraction_and_cognitive_machine_learning_techniques_for_recognition_of_lung_cancer_cells
21. Khan SA, Hussain S, Yang S, Iqbal K. Effective and reliable framework for lung nodules detection from CT scan images. *Scientific Reports*. 2019, 9(1), 4989. DOI:10.1038/s41598-019-41510-9.