Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

A Hybrid Approach to Feature Extraction and Information Gain-Based Reduction for Image Classification

Purushottam Das *

*Corresponding author, Department of Computer Sc. & Engineering, Graphic Era University, Dehradun, India. E-mail: pdas.nvs@gmail.com

Dinesh C. Dobhal

Prof., Department of Computer Sc. & Engineering, Graphic Era University, Dehradun, India. E-mail: dineshdobhal@gmail.com

Journal of Information Technology Management, 2025, Vol. 17, Special Issue, pp. 1-15
Published by the University of Tehran, College of Management
doi: https://doi.org/10.22059/jitm.2025.102918
Article Type: Research Paper

Received: January 17, 2025
Received in revised form: March 03, 2025
Accepted: June 13, 2025
Published online: August 01, 2025

Abstract

Image classification is a significant process in the field of computer science. It has applications in every field, such as spam detection in emails, medical diagnosis, image recognition, sentiment analysis, object detection, weather forecasting, pattern recognition, and security. Image classification deals with the grouping of images based on labels or characteristics. Feature extraction, feature selection, feature reduction, and classification are the main steps used to classify images. A medicinal and non-medicinal flowers data set is prepared by clicking images for the study. Methodology is used to achieve satisfactory classification results on the seeds, Wisconsin Diagnostic Breast Cancer, Heart Failure Clinical Records, and Wisconsin Prognostic Breast Cancer data sets, which are taken from the University of California, Irvine (UCI) repository. The proposed methodology suggests an efficient feature extraction and selection approach for data sets under consideration. An information gain-based genetic algorithm is used for feature reduction. It is performed on the extracted features to retrieve an optimized feature set. Fitness of the features is evaluated to choose the most relevant features. A neural network is used to classify the obtained feature subset. Better classification results are attained with the help of feature extraction and feature reduction.

Keywords: Image classification, feature extraction, feature reduction, information gain, UCI, ge-netic algorithm.

Introduction

Classification is a basic technique used to organize data into predefined categories. It includes models that can estimate the class of the given data based on its features. These techniques can be categorized as supervised, unsupervised, and semi-supervised methods. Supervised learning includes the classification of labeled datasets. It understands the relationship between input features and their labels to categorize test data accurately. Unsupervised classification deals with the categorization of input features without using any predefined labels. Labels are not present in unsupervised classification. It is grouping based on similar characteristics. Some techniques are clustering, like k-means, hierarchical, and Gaussian mixture models (Caldeira et al., 2020). Image classification deals with the grouping of images. It aims to teach machines to perceive and interpret visual data like humans. It includes feature extraction, pre-processing, image acquisition, feature selection, and classification. Image acquisition refers to the acquisition of images. Pre-processing involves resizing, normalizing, and augmenting to improve classification and classification performance. Feature extraction deals with retrieving features from the images. Various techniques are used to identify distinct characteristics of objects such as edges, shapes, colors, and textures. Image classification has applications in many fields such as medical imaging, automotive industry, agriculture, social media platforms, security and surveillance, deforestation, gaming industry, education, and manufacturing industry (Agrawal & Bhatnagar, 2023; Singh & Kumar, 2023; Wang et al., 2019).

Feature selection is a vital process in machine learning. It involves identifying the most relevant features in a dataset to improve the performance of the model. It focuses on reducing the dimensionality of the dataset by choosing features that contribute the most to classification. There are many irrelevant and redundant features, which do not contribute much to the classification and should be removed. This increases the computational efficiency. Feature selection becomes more significant when we are dealing with a large dataset. Three types of feature selection are filter, wrapper, and embedded methods (Kwak & Choi, 2002). Filter methods are based on the calculation of the relationship between each input feature and the target variable. Features having high correlation with the target variable are considered more relevant. These methods yield results in less time and work independently of any particular machine learning algorithm. Wrapper methods assess subsets of features by training and testing a model repeatedly. These methods assess the efficiency of an algorithm on various feature subsets and choose the combination that optimizes the predictive accuracy of the model. Common techniques used here are backward elimination, forward selection, and recursive feature elimination. Wrapper methods yield better accuracy by considering feature interactions, but they are computationally expensive for large datasets. Embedded methods obey the hybrid approach that uses the concept of both earlier approaches. These methods automatically identify the most important features as part of the algorithm's operation. A few examples are decision trees, LASSO (least absolute shrinkage and selection operator), and tree-based ensemble methods like gradient boosting and random forests. These methods tend to provide a balance between filter and wrapper approaches. They are computationally efficient and capable of handling feature interactions (Dhal & Azad, 2022; Agrawal et al., 2021; Al-Tashi et al., 2020).

Beyond traditional techniques, the latest techniques, such as evolutionary algorithms and ensemble methods, such as particle swarm optimization and genetic algorithms, are gaining popularity. They simulate the natural selection process to iteratively find an optimal feature subset. Genetic algorithms are encouraged by the process of natural selection. It was given by John Holland in the 1970s. It involves three main genetic operators: selection, crossover, and mutation (Katoch et al., 2021). Mutual information is a measure derived from information theory that quantifies the statistical dependency between random variables. Mutual information can significantly increase the performance of a genetic algorithm when incorporated with it. One primary application is in feature selection and dimensionality reduction in high-dimensional optimization problems. It evaluates the mutual information between variables to identify and keep the most informative features while discarding redundant and irrelevant features (Battiti, 1994). This reduces the search space quite significantly. After retrieving an optimized set, the next step is to apply classification techniques to get the results. Neural networks are used primarily for this. It includes an input layer, output layers, and hidden layers. The input layer accepts the data. The hidden layer achieves computations through weighted connections and activation functions, i.e., ReLU, sigmoid, etc. The output layer displays the predictions or classification results. Data is passed through layers where each neuron applies a weighted sum and activation function. Errors between predictions and true labels are calculated using a loss function. Weights are updated via backpropagation and optimization (gradient descent) to minimize the error (Li et al., 2021). Neural networks have applications in several fields, like natural language processing, image recognition, speech processing, and more. Neural networks handle complex, non-linear relationships and large datasets effectively (Fan et al., 2021).

A detailed survey of image classification algorithms, feature selection and optimization methods, and classification techniques such as neural networks and support vector machines, is performed. Fatima and Pasha gave a detailed analysis of various machine learning techniques. A detailed survey on machine learning methods used for the estimation of hepatitis, dengue, liver, diabetes, and heart diseases is given (Fatima & Pasha, 2017). Gambella et al. conducted a study about machine learning techniques and optimization algorithms (Gambella et al., 2021). Various mathematical models are studied and proposed for deep learning, clustering, classification, and regression. Xue et al. presented a detailed and comprehensive study of the evolutionary algorithms used in feature set optimization. The contribution of different techniques is weighed in using a feature selection process (Xue et al., 2015). Present challenges are explored along with finding scope for future research. Saranya and Pravin proposed a review of feature selection methods used for the detection of diseases.

Feature selection algorithms confront significant challenges regarding efficiency and efficacy due to the current rise in data variety and speed (Saranya & Pravin, 2021). Thejas et al. proposed a new hybrid method on the basis of mini-batch normalized mutual information, which integrates both filter and wrapper approaches. This hybrid model is a two-step structure. These steps consist of ranking the features first, and then feature subset selection is done on the basis of the ranks of the features (Thejas et al., 2019). Tali et al. presented a biodiverse database on medicinal plants of Jammu and Kashmir state. It provides an extensive ethno-medicinal review done over the last 10 years. The database has a total of 1,123 plants and their medicinal uses. This survey discovered 20 important medicinal species used to recover from around 25 illnesses. Mostly used plants are Acotinum heterophyllum and Taraxacum officinale (Tali et al., 2019). Panyadee et al. presented a study that aims to find the drivers of diversity and medicinal plants used in home gardens. A comparison is made between four traditional groups of northern Thailand (Thai Lisu, Karen, Lahu, and Yuan). These groups are interviewed about the prominently used medicinal plants, their primary and secondary benefits. 95 medicinal plant species are recorded from four home gardens (Panyadee et al., 2019).

Singh et al. proposed an algorithm for an image segmentation approach that can be applied to detect diseases automatically, along with classifying them. A review on various classification approaches, which can be utilized for disease detection using plant leaves, is presented (Singh et al., 2015). Thamilselvan and J. Sathiaseelan discussed the importance of data mining, image mining, and classification. They compared some significant algorithms used for data mining, like artificial neural network, CART, kNN, SVM, etc. (Thamilselvan & Sathiaseelan, 2015). Sun et al. proposed a new method based on genetic algorithms for developing the architecture and connection weight initialization of a deep CNN. An effective dynamic-length gene encoding method is created for displaying the various building blocks and depth in a CNN (Sun et al., 2019). Jha and Dutta developed a module for the estimation of leukemia, based on deep learning, from blood smear images. This technique gave a 98.7% classification performance (Jha & Dutta, 2019). Khan et al. proposed an improved form of a texture-based feature descriptor using local tetra pattern (LTrP), known as the uniform variant of LTrP (ULTrP). For future scope, a hybrid feature selection process can be used for efficient feature selection (Khan et al., 2022). Sachar and Kumar suggested a recent review on various leaf extraction approaches. Medicines obtained from plants have no side effects and are accessible at lower costs (Sachar & Kumar, 2021).

Methodology

The proposed methodology is a hybrid approach that has steps such as image data set preparation, feature extraction and selection, feature reduction, and classification. Given steps should be followed to achieve the proposed Information gain-based hybrid approach:

(a) Data set collection- An Image dataset is prepared by clicking images of various medicinal and non-medicinal flowers;

(b) Feature extraction and selection- Features are extracted from the images under consideration. Common 15 features are selected using the stated techniques.

(c) Feature reduction: An information gain-based genetic algorithm is applied to get for feature subset. Fitness of each feature is measured to check and appropriate features are selected, which contribute the most in feature selection;

(d) Neural network: A Neural network is used to categorize the improved feature dataset. Hidden layers, neurons, training, testing, and validation parameters are adjusted to get efficient classification accuracy.

(e) Classification performance: A comparative analysis is done on attained classification results of various data sets using the proposed approach.

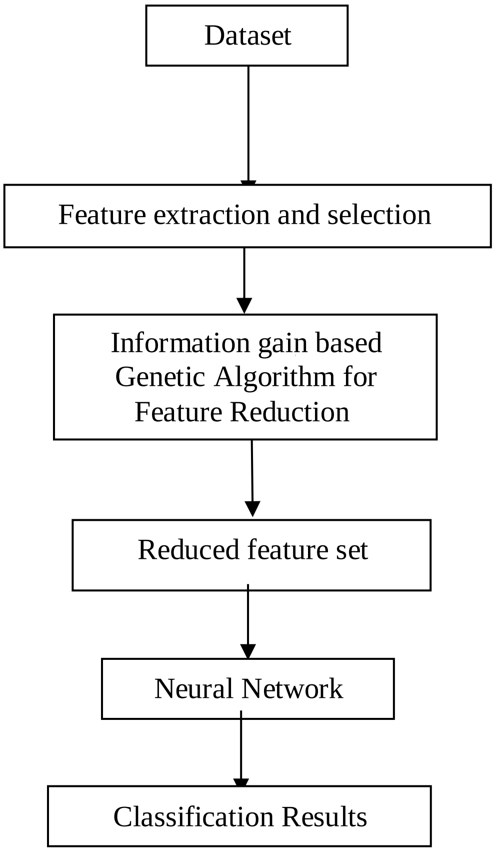

Dataset
Feature extraction and selection
Information gain based Genetic Algorithm for Feature Reduction
Reduced feature set
Neural Network
Classification Results

Figure 1. Framework of the proposed hybrid feature set optimization and classification approach

The framework, as shown in Figure 1, includes steps of data set preparation, feature extraction, feature selection, feature reduction, and classification of the optimized subset by a neural network.

Feature extraction and selection consist of a function that is used to excerpt features based on color from the given flower images. It finds common features between the two images. OpenCV is used to load images, and then the images are converted to the HSV color space. Flower regions are extracted using color thresholds. The best 15 common features are selected with their values. The given steps are used to extract color histograms of flowers from the two images taken. It is used to identify common feature indices and extract the top 15 common features. These common features are saved in a file along with their feature values. The algorithm for this is given here: let us take two images, and their path is image1_path and image2_path. Two functions are used here: extract_flower_features(image_path) and find_common_features(image1_path, image2_path).

```
Function Definition - extract_flower_features(image_path):

image←cv2.imread(image_path)

hsv←cv2.cvtColor(image, cv2.COLOR_BGR2HSV)

lower_hue←[0,50,50],upper_hue←[30,255,255]

mask←cv2.inRange(hsv, lower_hue, upper_hue)

hist←cv2.calcHist([image], [0, 1, 2], mask, [8, 8, 8], [0, 256, 0, 256, 0, 256])

hist←hist.flatten()

Return the histogram

Function Definition - find_common_features(image1_path, image2_path)

features1←extract_flower_features(image1_path)

features2←extract_flower_features(image2_path)

common_indices←set(range(len(features1)))∩set(range(len(features2)))

sorted_common_indices←sorted(common_indices, key=lambda x: features1[x] + features2[x], reverse=True)

top_common_indices←sorted_common_indices[:15]

results_df←pd.DataFrame('Feature_Index': top_common_indices,

'Feature_Value_Image1': [features1[i] for i in top_common_indices],

\text{'Feature\_Value\_Image2': [features2[i] \text{ for } i \text{ in } top\_common\_indices]})}
```

results_df.to_csv('common_flower_features.csv', index=False)

"Common flower features saved to common_flower_features.csv"

Main Program Execution:

image1_path←"N11.jpg", image2_path←"N21.jpg"

Call the function find_common_features(image1_path, image2_path)

A subset of features is selected using a genetic algorithm-based approach. A few sub-functions are also used in this section. Firstly, a function is used to calculate the entropy of the feature set. Numbers of data values of each cell are used to determine the probability and entropy using H=−∑p.log(p)H, where p denotes the probability of data points from every grid cell. H is the output that represents the final entropy value. Secondly, a genetic algorithm is used to select a feature subset with reference to target values. Fitness of each feature set is measured to balance redundancy and relevance. Based on fitness values, the population is ranked, and feature sets with good performance are selected for crossover. Thirdly, a function is used to calculate statistical parameters such as entropy and mutual information for the input feature set and target feature set. Following notations are used in this section: p(a) and p(b): probability mass functions; H(A): Entropy of a discrete random variable A; $H(A|B)$: Conditional entropy for continuous random variables; $I(A;B)$: The information gain among two continuous random variables A, B with joint pdf $p(a,b)$ (Zhou et al., 2022). Information gain is derived as follows:

$$H(A) = -\sum_{x\in Z} p(a) \log p(a)$$

$$H(A|B) = -\iint p(a,b) logp(a|b)\, da\, db$$

$$H(B|A) = -\sum_{a\in Z}\sum_{b\in B} p(a,b) \log p(b|a)$$

$$H(A,B) = -\sum_{a\in Z}\sum_{b\in B} p(a,b) \log p(a,b)$$

$$= H(A,B) = H(A) + H(B|A)$$

$$I(A;B) = \sum_{a,b} p(a,b) log \frac{p(a,b)}{p(a)p(b)}$$

$$= H(A) - H(A|B)$$

$$I(A;B) = H(B) - H(B|A)$$

Since $H(A,B) = H(A) + H(B|A)$as shown, we have…

$$I(A;B) = H(A) + H(B) - H(A,B)$$

$$I(A;A) = H(A) - H(A|A) = H(A)$$

Mutual information and entropy:

$$I(A;B) = H(A) - H(A|B)$$

$I(A;B) = H(B) + H(B|A)$

$I(A;B) = H(A) + H(B) + H(A,B)$

$I(A;B) = I(B;A)$

$I(A;A) = H(A)$

The conditional information gain is:

$I(A;B|Z) = H(A|Z) - H(A|B,Z)$

$= E_{p(a,b,z)} log \frac{p(A,B|Z)}{p(A|Z)p(B|Z)}$

Genetic algorithm uses the information gain-based fitness function: $\mathrm{i\_MIrs} - \mathrm{t} * \mathrm{i\_MIrr}$, where, mean of information gain among output and input features is denoted by i_MIrs and the mean of information gain among input features is denoted by i_MIrr. For classification, a function is used to create a pattern recognition system with a specified number of neurons. Data dataset is divided into training, testing, and validation subsets. The neural network is trained using training data, and its efficiency is measured using the predicted and actual target outputs. Confusion matrix and ROC plots are used to display the classification results. Network structure is also represented, and the trained network is saved into a file. The final output consists of classification results and performance metrics of the resultant model.

Results

Five types of data sets are taken into consideration for achieving classification results on:

(a) Medicinal and non-medicinal flowers data set:

Two classes of medicinal flowers are nasturtium and Tagetes erecta. Moss-rose purslane and Surfinia are the two classes of non-medicinal flowers. A total of four classes are under study; each class has 20 images and 19 unique comparisons. Each comparison has two unique columns, which gives us 38 columns for a class. So, it means each class has 38 instances. The total of four classes has 152 instances. A set of 15 common features is selected using efficient feature extraction and selection techniques. The top 10 significant attributes are considered for dimensionality reduction using a mutual information-based genetic algorithm approach. A feature subset of 4 features is used for achieving classification results.

(b) Seed data set:

Seed data consists of three classes, and each class has 70 instances. There are a total of 210 instances. A total of seven features are: 1. length of kernel groove, 2. asymmetry coefficient, 3. width of kernel, 4. length of kernel, 5. compactness C = 4*pi*A/P^2, 6. perimeter P, 7. Area A (Charytanowicz et al., 2010).

(c) Wisconsin Diagnostic Breast Cancer (WDBC) data set:

WDBC contains a total of 569 instances, 2 classes, and a total of 32 attributes. Given 10 features are estimated: a) fractal dimension, b) symmetry, c) concave points, d) concavity, e) compactness, f) smoothness, g) area, h) perimeter, i) texture, j) radius (Wolberg, 1990).

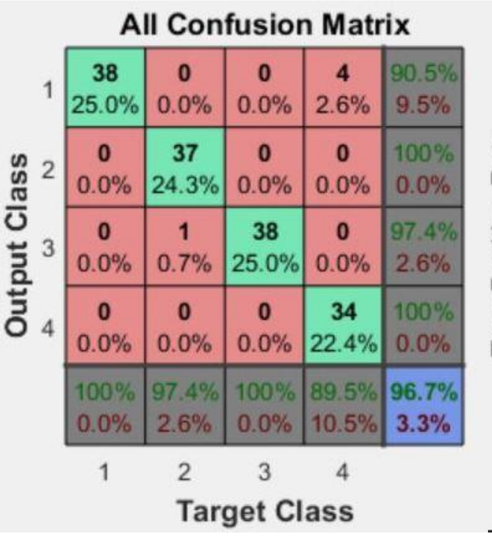
(d) Wisconsin Prognostic Breast Cancer (WPBC) data set:

198 instances, 2 classes, and 34 features are present in the WPBC dataset. There are a total of 34 features, but the prime 10 features are: a) fractal dimension, b) symmetry, c) concave points, d) concavity, e) compactness, f) smoothness, g) area, h) perimeter, i) texture, j) radius (Wolberg, 1990).

(e) Heart failure clinical records (HFCR) data set

In the HFCR dataset, 299 patients have heart problems. There are two classes and a total of 13 features, out of which 12 are input features and 1 is an output feature. Deceased is the output feature, and Age, Diabetes, Time, Serum creatinine, Creatinine phosphokinase, Anaemia, Smoking, Platelets, Gender, High blood pressure, Ejection fraction, and Serum Sodium are 12 input features (Heart Dataset, 2020).

The above section is displaying a confusion matrix and ROC plots diagram as classification results for medicinal and non-medicinal data sets are in Figure 2a, 2b:



All Confusion Matrix

Output Class	1	2	3	4	
1	38 25.0%	0 0.0%	0 0.0%	4 2.6%	90.5% 9.5%
2	0 0.0%	37 24.3%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	1 0.7%	38 25.0%	0 0.0%	97.4% 2.6%
4	0 0.0%	0 0.0%	0 0.0%	34 22.4%	100% 0.0%
	100% 0.0%	97.4% 2.6%	100% 0.0%	89.5% 10.5%	96.7% 3.3%

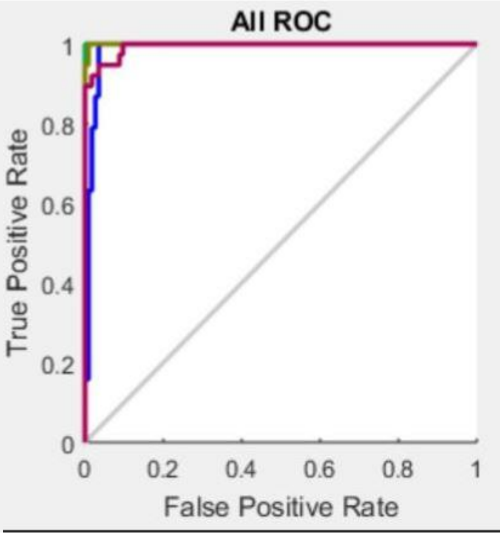
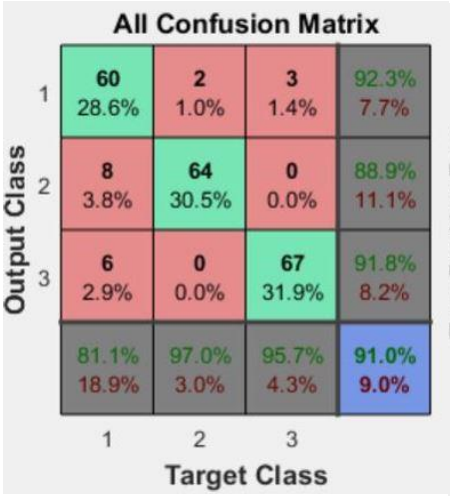


Figure 2a. Confusion matrix **Figure 2b. ROC plot**

The following section displays a confusion matrix and ROC plots as classification results for seed, wdbc, and wpbc are shown in Figure 3a, 3b, 4a, 4b, 5a, 5b.



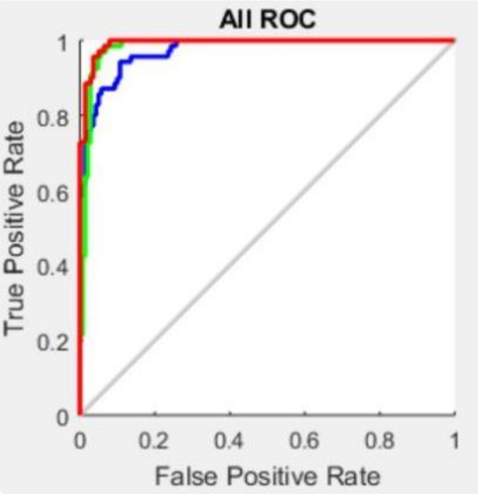
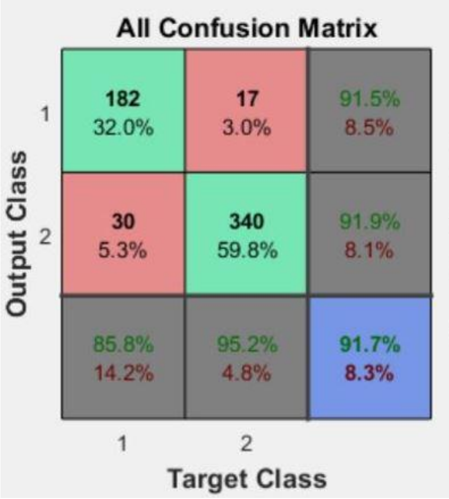


Figure 3a. Confusion matrix

Figure 3b. ROC plot



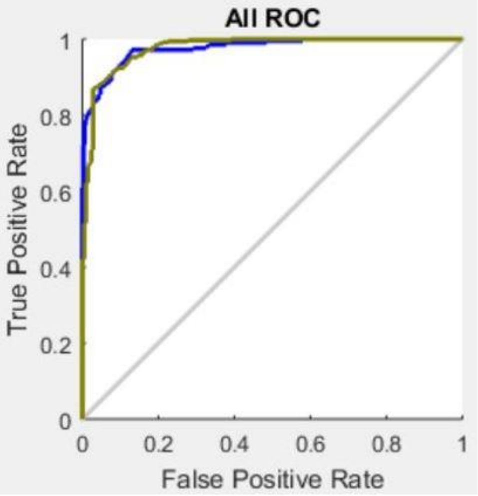
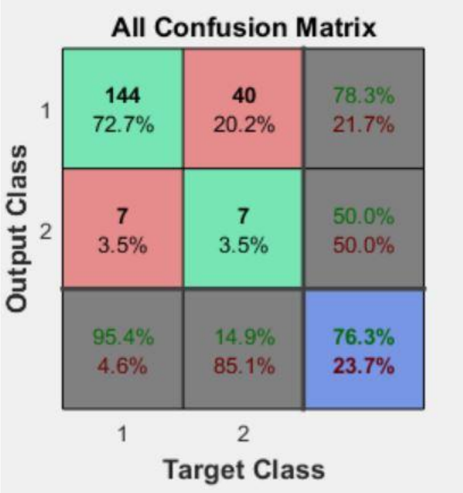


Figure 4a. Confusion matrix

Figure 4b. ROC plot



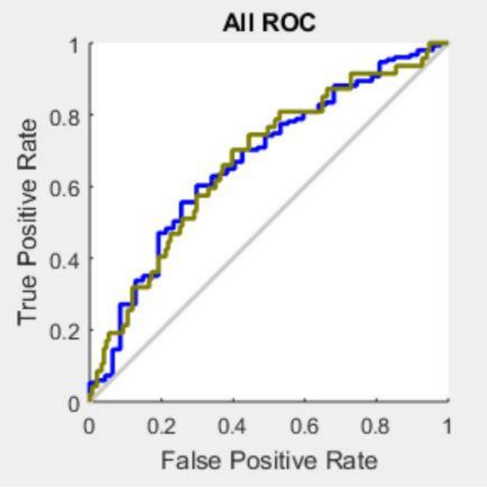


Figure 5a. Confusion matrix

Figure 5b. ROC plot

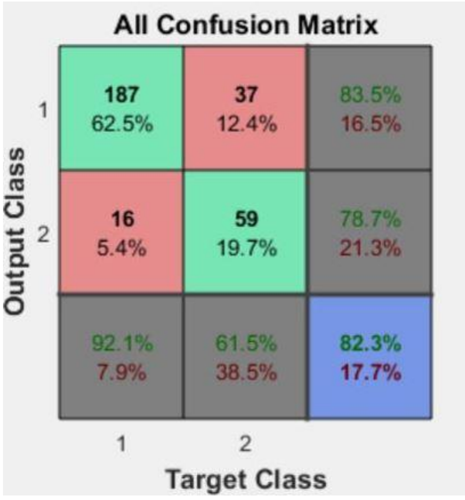


Figure 6a. Confusion matrix

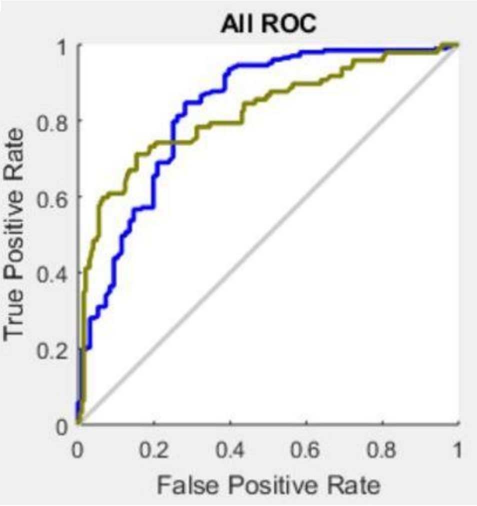


Figure 6b. ROC plot

The above section displays the confusion matrix and ROC plots as classification accuracy for the HFCR dataset in Figures 6a and 6b.

Table 1. Classification Performance on various datasets under consideration

Data	Feature Count	Selected feature-set	Results			
			Training	Testing	Validation	Total
Medicinal and Non-medicinal Flowers	4	(5, 3, 9, 2)	98.1	100.00	87.0	96.7
Seed	4	(2, 3, 6, 1)	92.5	81.3	93.8	91.0
WDBC	2	(2, 5)	90.7	92.9	95.3	91.7
WPBC	4	(33, 15, 16, 31)	77.5	70.0	76.7	76.3
HFCR	3	(10, 2, 12)	84.7	77.8	75.6	82.3

Classification performance achieved on various datasets under consideration is given in Table I.

Discussion

We have used four standard datasets, i.e., Seeds, Wisconsin Diagnostic Breast Cancer, Heart Failure Clinical Records, and Wisconsin Prognostic Breast Cancer dataset, and one in-house dataset, i.e., medicinal and non-medicinal datasets, which was achieved by clicking images. Features are extracted from the image dataset, and 15 common features are selected. An information gain-based genetic algorithm approach is used for attaining an improved feature subset. Classification results are improved as we have reduced the feature space significantly by eliminating irrelevant and redundant attributes. Comprehensive exploration is not suggested in the situation of a huge feature dataset. Thus, the proposed framework is more efficient in the case of higher data dimensionality.

The Seeds dataset has three classes, 7 attributes, and 210 instances. Four features (2, 3, 6, 1) are taken after feature reduction. The WDBC dataset has 12 features and 569 instances; WPBC consists of 34 attributes and 198 instances. In WDBC and WPBC, features (2, 5) and (33, 15, 16, 31) are chosen respectively. Features (12, 2, 10) are selected for the HFCR dataset. Classification results for Seeds, Wisconsin Diagnostic Breast Cancer, Heart Failure Clinical Records, and Wisconsin Prognostic Breast Cancer datasets are 91.0%, 91.7%, 76.3%, and 82.3% respectively. A medicinal and non-medicinal flowers dataset is prepared. The flowers dataset consists of 20 images per class, which contains 19 unique comparisons. Each comparison has 2 columns, so 38 instances are in a single class. There are 4 classes, having a total of 80 images and 152 instances. 10 attributes are extracted and selected using an efficient feature extraction technique. Classification accuracy on medicinal and non-medicinal flowers is achieved as 96.7%.

Conclusion

A hybrid feature extraction and information gain-based feature reduction approach is proposed in this paper. Classification performance is analyzed on standard datasets, namely Seeds, Wisconsin Diagnostic Breast Cancer, Heart Failure Clinical Records, and Wisconsin Prognostic Breast Cancer datasets, which are attained from UCI. A medicinal and non-medicinal flowers dataset is prepared by taking images in different scenarios. An information gain-based feature reduction approach is used to evaluate the fitness of features. Strongly relevant features are chosen using an information gain-based genetic algorithm. Satisfactory classification results are attained on all the datasets. 91.0% and 76.3% classification accuracies are retrieved on Seeds and Heart datasets, respectively. Classification performance on WDBC and WPBC is achieved as 91.7% and 82.3%. Training, testing, validation, and total classification performances are 98.1%, 100%, 87.0%, and 96.7%, respectively, on the medicinal and non-medicinal flowers dataset, choosing (5, 3, 9, 2) as the most significant features. For future scope, medicinal and non-medicinal datasets can be expanded, and more varieties and images can be added. Further, other techniques of feature optimization, like particle swarm optimization, can be explored on the datasets under study.

Acknowledgement

I sincerely thank my supervisor, Dr. Dinesh C. Dobhal, Professor in the Department of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, for their continuous guidance and immense support. All the related and referenced articles are cited and acknowledged appropriately. I thank my friends for helping to overcome the difficulties I faced while writing the manuscript.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). *Ieee Access*, *9*, 26766-26791.

Agrawal, K., & Bhatnagar, C. (2023). F-mim: Feature-based masking iterative method to generate the adversarial images against the face recognition systems. *Journal of Information Technology Management*, *15*(Special Issue: EIntelligent and Security for Communication, Computing Application (ISCCA-2022)), 80-93.

Al-Tashi, Q., Abdulkadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2020). Approaches to multi-objective feature selection: a systematic literature review. *IEEE Access*, *8*, 125076-125096.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537-550.

Caldeira, M., Martins, P., Costa, R. L. C., & Furtado, P. (2020). Image classification benchmark (ICB). *Expert Systems with Applications*, *142*, 112998.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P., & Lukasik, S., (2010). "Seeds [Dataset]". *UCI Machine Learning Repository.* https://doi.org/10.24432/C5H30K.

Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, *52*(4), 4543-4581.

Fan, F. L., Xiong, J., Li, M., & Wang, G. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *5*(6), 741-760.

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1-16.

Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2021). Optimization problems for machine learning: A survey. *European Journal of Operational Research*, *290*(3), 807-828.

Heart Dataset, (2020). Heart Failure Clinical Records [Dataset], *UCI Machine Learning Repository.* https://doi.org/10.24432/C5Z89R.

Jha, K. K., & Dutta, H. S. (2019). Mutual information based hybrid model and deep learning for acute lymphocytic leukemia detection in single cell blood smear images. *Computer methods and programs in biomedicine*, *179*, 104987.

Katoch, S., Chauhan, S. S., & Kumar, V., (2021). A review on genetic algorithm: past, present, and future. *Multimedia tools and applications", 80*, 8091-8126.

Khan, A. H., Sarkar, S. S., Mali, K., & Sarkar, R. (2022). A genetic algorithm based feature selection approach for microstructural image classification. *Experimental Techniques*, 1-13.

Kwak, N., & Choi, C. H. (2002). Input feature selection for classification problems. *IEEE transactions on neural networks*, *13*(1), 143-159.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, *33*(12), 6999-7019.

Panyadee, P., Balslev, H., Wangpakapattanawong, P., & Inta, A. (2019). Medicinal plants in homegardens of four ethnic groups in Thailand. *Journal of ethnopharmacology*, *239*, 111927.

Sachar, S., & Kumar, A. (2021). Survey of feature extraction and classification techniques to identify plant through leaves. *Expert Systems with Applications*, *167*, 114181.

Saranya, G., & Pravin, A. (2021). Feature selection techniques for disease diagnosis system: A survey. In *Artificial Intelligence Techniques for Advanced Computing Applications: Proceedings of ICACT 2020* (pp. 249-258). Springer Singapore.

Singh, M. K., & Kumar, A. (2023). Cucumber leaf disease detection and classification using a deep convolutional neural network. *Journal of Information Technology Management*, *15*(Special Issue: EIntelligent and Security for Communication, Computing Application (ISCCA-2022)), 94-110.

Singh, V., & Misra, A. K. (2015). Detection of unhealthy region of plant leaves using image processing and genetic algorithm. In *2015 International Conference on Advances in Computer Engineering and Applications* (pp. 1028-1032). IEEE.

Sun, Y., Xue, B., Zhang, M., & Yen, G. G. (2019). Evolving deep convolutional neural networks for image classification. *IEEE Transactions on Evolutionary Computation*, *24*(2), 394-407.

Tali, B. A., Khuroo, A. A., Ganie, A. H., & Nawchoo, I. A. (2019). Diversity, distribution and traditional uses of medicinal plants in Jammu and Kashmir (J&K) state of Indian Himalayas. *Journal of Herbal Medicine*, *17*, 100280.

Thamilselvan, P., & Sathiaseelan, J. (2015). A comparative study of data mining algorithms for image classification. *Int. J. Educ. Manage. Eng*, *5*(2), 1-9.

Thejas, G. S., Joshi, S. R., Iyengar, S. S., Sunitha, N. R., & Badrinath, P. (2019). Mini-batch normalized mutual information: A hybrid feature selection method. *IEEE Access*, *7*, 116875-116885.

Wang, Y., & Wang, Z. (2019). A survey of recent work on fine-grained image classification techniques. *Journal of Visual Communication and Image Representation*, *59*, 210-214.

Wolberg, W., (1990). “Breast Cancer Wisconsin (Original) [Dataset]”. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5HP4Z.

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, *20*(4), 606-626.

Zhou, H., Wang, X., & Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. *Applied intelligence*, *52*(5), 5457-5474.

Bibliographic information of this paper for citing:

Das, Purushottam & Dobhal, Dinesh C. (2025). A Hybrid Approach to Feature Extraction and Information Gain-Based Reduction for Image Classification. *Journal of Information Technology Management*, 17 (Special Issue), 1-15. https://doi.org/10.22059/jitm.2025.102918