MACHINE LEARNING-BASED CLASSIFICATION OF INDIAN CASTE CERTIFICATES USING GLCM FEATURES

Subramani C [1], SureshR [2], Muralidhara B L [1]

[1] Department of Computer Science and Applications,
Bangalore University, India
[2] Department of Statistics, Bangalore University, Bengalore, India

ABSTRACT

In today's world, there is a growing prevalence of fraudulent Caste certificate schemes, posing a threat to society. It is essential to prevent the issuance of these false documents as they can potentially disrupt the social order. Accurate classification and verification of Indian Caste certificates to prevent counterfeiting can ensure equitable access to benefits. Implementing such digital systems can help thwart these scams, creating a community devoid of counterfeit documents and promoting a strong social welfare framework. The detection process includes analyzing scanned copies against authentic references using image processing methods. This study tackles the issue using texture features extracted through the Gray Level Co-occurrence Matrix (GLCM) and diverse machine learning (ML) algorithms. The approach encompasses image pre-processing, GLCM feature extraction, and the application of classifiers such as K-Nearest Neighbor (KNN), Decision Tree (DT),Support Vector Machine (SVM),Naive Bayes (NB), andRandom Forest (RF). Assessments based on accuracy, confusion matrices, and AUC (area under curve) scores indicate that the Naive Bayes classifier outperforms other methods with 100%accuracy and a robust AUC score. These findings indicate that integrating GLCM features with ML algorithms offers a dependable solution for Caste certificate authentication.

KEYWORDS

Image processing, Classification, GrayLevel Co-occurrence Matrix, Machine learning.

1. INTRODUCTION

In India, Caste certificates confirm eligibility for social, educational, and economic benefits, ensuring access to reservations in education, government employment, and welfare programs. Nevertheless, extensive forgery undermines these systems, resulting in the unjust distribution of benefits and placing genuine beneficiaries at a disadvantage.

Authenticating Caste certificates is crucial given their pivotal role. Conventional manual verification methods are both time-consuming and prone to errors. Image classification presents a promising solution that automates the verification process. Sophisticated image-processing methods can discern between genuine and counterfeit certificates, thereby guaranteeing fair allocation of benefits. This study addresses the task of accurately classifying Indian Caste certificate images, emphasizing the inefficacy of manual verification. The paper aims to devise an efficient method utilizing GLCM features and ML algorithms, harnessing texture analysis to establish a robust verification solution.

DOI:10.5121/cseij.2025.15102

The field of image classification has attracted considerable attention due to itsinfluential advancements and successful implementations in domains such as biomedicine, remote sensing, industry, and many more [1]. In supervised machine learning image classification, two pivotal processes come into play training and testing. During the training phase, images with known categories undergo analysis using feature extraction methods to pinpoint significant features, which are stored as feature vectors. In the testing phase, a new image is introduced to the system to forecast its class [2]. Feature extraction is vital, as the machine learning algorithm's performance relies on these features [3].

This research uses theGLCM to derive texture features from Caste certificate images, calculating energy, correlation, dissimilarity, homogeneity, and contrast across various distances and angles. These GLCM features will be categorized using ML algorithms. The performance will be assessed based on accuracy, confusion matrices, and AUC scores. The findings will be scrutinized to compare classifier performance and highlight their strengths and weaknesses. Since it was first proposed by Haralick in 1973, the GLCM has been utilized as a texture analysis method to characterize spatial patterns in images, especially in Gray-level photomicrographs [4]. GLCM has gained widespread acceptance for image classification through the use of second-order statistical measures [5]. It serves as a powerful texture descriptor, providing high accuracy and computational efficiency in comparison to alternative texture extraction techniques [6].

Our proposed work is organized as follows: related work is covered in Section 2, the dataset and methodology are outlined in Section 3, findings and analysis are discussed in Section 4, and the study is concluded in Section 5.

2. LITERATURE REVIEW

Research in computer vision has extensively focused on image classification, often entailing the extraction of handcrafted features and the use of classifiers. Although GLCM features and MLfind applications in diverse fields, their potential in verifying Caste certificates has not been thoroughly explored. This study seeks to address this gap by showcasing the effectiveness of GLCM features in the classification of Caste certificate images.

The statistical texture analysis method known as the GLCM takes into account pixel spatial relationships. GLCM features offer valuable texture information, including energy, correlation, dissimilarity, homogeneity, and contrast. Despite its proven efficacy in various classification tasks, GLCM's utilization in document verification, particularly for Caste certificates, remains limited. While ML algorithms like Decision Tree, RF,KNN, SVM,and Naive Bayes are well-suited for image classification, their application to Caste certificate classification using GLCM features lacks comprehensive documentation.

In research work [7], there is a critical necessity to identify image manipulation, especially considering the widespread use of images as documentary proof in forensic investigations and various other fields. The objective of image fraud detection based on pixels is to verify digital images without requiring prior information about the original image. Images can undergo tampering through several methods including splicing,resampling, copy-moveand the addition or removal of objects. Additionally, as mentioned in the work [8], visually detecting image alterations is extremely challenging for humans. The occurrence of digitally manipulated forgeries in mainstream media and on the internet is expanding quickly. According to Hany Farid, it was noted that digital forgeries, despite lacking visual cues, have the potential to modify an image's underlying statistics. Image forensic tools can be categorized into five primary types: techniques based on pixels, based on formats, based on cameras, based on physical properties, and techniques based on geometric features[9].

Hayat, Khizar, and Tanzeela Qazi. introduced a forgery detection technique that involves combiningdiscrete cosine transform (DCT)and discrete wavelet transform (DWT) to reduce features. In this method, DCT is utilized on blocks generated from the DWT-processed image, and compared using correlation coefficients. Additionally, as part of the experiments to assess the detection approach, a mask-based tampering method was formulated and tested. [10]. In their research, the authors employed an Artificial Neural Network (ANN) classifier to distinguish tomatoandcotton leaf diseases based on features such as standard deviation, contrast, homogeneity, energy, mean, correlation, and variance, attaining an overall accuracy of 92.5% [11]. Another investigation by [12] utilized GLCM texture features in combination with an SVM classifier to categorize infected tomato leaves, achieving a superior accuracy of 99.83% by utilizing a linear kernel function. In an investigation conducted by [13], the classification of sunflower crop diseases, was explored using Multi-Class SVM, NB,KNN,and Multinomial Logistic Regression (MLR). The classification models employed color and texture features such asenergy, standard deviation,coarseness, mean, contrast, andhomogeneity. MLR yielded the best average accuracy of 92.57%, with all classifiers achieving 100% accuracy for healthy leaves.

In their work, A. Singh, Aditi, and Harjeet Kaur. introduced a multi-layer SVM approach with a linear kernel function to categorize potato leaf diseasesbased on GLCM texture features. The SVM classifier attained an overall accuracy of 95.99%, along withrecall, precision, and F1-scores of 96.12%, 96.25%, and 96.16%, respectively [14]. In the study conducted by Naveed Iqbal GLCM features were extracted from grayscale photos collected by a drone. ML techniques such as RF, NB, SVM, and Neural Network (NN)were employed to classify various crop types. The findings show that ML algorithms exhibited notably superior performance when utilizing GLCM features, leading to an overall accuracy enhancement of 13.65% [15].In their work, Mireille Pouyap et al. suggested the utilization of the GLCM for performing texture analysis on vibration signals within images. They combined PCA and Sequential Features Extraction (SFE) methods to select pertinent features. When tested with a multiclass-Naive Bayes classifier, this approach achieved a success rate of 98.27%, displaying enhanced efficiency and promising outcomes in comparison to existing methods [16]. Li, Dian, Cheng Wu, and Yiming Wang.suggest an anti-counterfeit method for iris detection using a binary classification neural network and an enhanced Modified-GLCM. This method outperforms the leading performance achieved in LivDet-Iris2017 and traditional texture analysis methods. Furthermore, the study assesses the potential risk of iris adversarial samples on the iris performance verification system through iris texture extraction [17].

The authors Padmavathi and Maya V. Karkigoal is to derive texture characteristics from brain tumor cases and categorize them as benign or malignant utilizing,classification phases, feature extraction, andsegmentation. K-means clustering is employed for segmentation and for selecting the region of interest. Textural information is collected through GLCM, HOG, and LBP patterns. The research assesses the accuracy of ANN, k-NN, and SVM classifiers in classifying tumors in brain MR images. The findings indicate that combining GLCM, LBP, and HOG feature extraction with an ANN using the ML training algorithm yields higher accuracy and superior performance in distinguishing benign and malignant tumors compared to other classifiers [18]. In their work, Barburiceanu, Stefania, Romulus Terebes, and Serban Meza [19] introducea technique that combines feature vectors from Local Binary Patterns (LBP) and GLCM methods. By employing classifiers such as SVM, k-NN, and RF, their approach surpasses traditional deep-learning networks and other customized texture feature extraction methods. Despite a modest number of images per class, the suggested method enhances discrimination capability and produces encouraging results. GLCM was utilized which was subsequently condensed to an optimal subset through PCA. The research revealed that the amalgamation of GLCM with PCA for feature reduction leads to elevated classification accuracy when employingANN for image categorization [20].

3. PROPOSED WORK

The study exclusively focuses on GLCM features, which capture important texture information. However, the incorporation of supplementary features or the utilization of alternative machine learning techniques has the potential to improve classification performance. The development of image forgery detection systems involves three primary stages: The proposed approach for our work is illustrated in Figure. 1.

1. Image pre-processing
2. Feature extraction using GLCM statistical features
3. Classification using various ML algorithms.

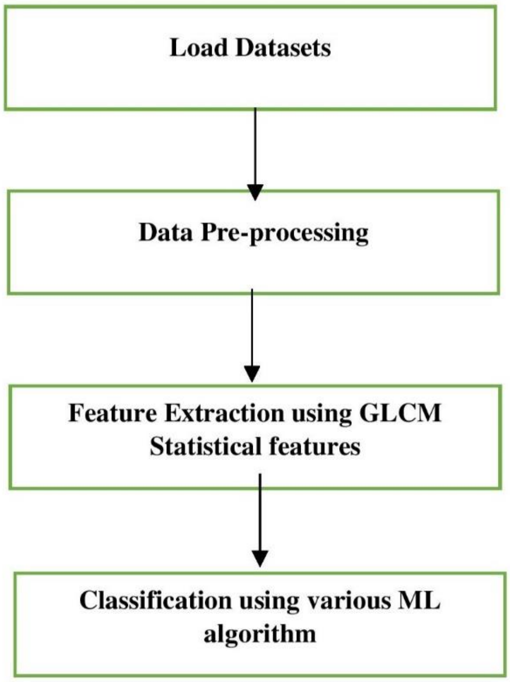


Figure1:Proposed flowchart for the model

3.1. Data pre-processing

3.1.1. Data Collection

We acquired a thousand images of Karnataka Scheduled Caste and Scheduled Tribe Caste Certificates from the Atalji Janasnehi Kendra, Nadakacheri Directorate, Karnataka Government. These images are in jpg format with a resolution of 300 dpi. Abnormal data samples were generated using Adobe Photoshop to evaluate the proposed ML models. Figure. 2. a and 2. b displays the samples of both original and fake samples respectively.

(a)

(b)

3.1.2. Image pre-Processing

The dataset was partitioned with 80% allocated for training and 20% reserved for testing, with the images resized to 400x400 pixels. Utilizing the original high-resolution images (1653x2339) in the ML model demands substantial computing resources and can lead to slower training, potentially resulting in the loss of crucial information. Pre-resizing enhances efficiency and effectiveness, accelerates computation time, and improves model stability. Rescaling pixel values between 0 and 1 is necessary for activation function requirements and faster convergence. Moreover, converting images to grayscale simplifies texture analysis.

3.2. GLCM- Based Feature Extraction

GLCM derives texture features from pre-processed images by determining the frequency of pixel pairs exhibiting specific values and spatial relationships. GLCM is computed for each image at various distances (1, 3, 5 pixels) and angles (0, 45, 90, 135 degrees). The extracted features include contrast, correlation, dissimilarity, energy, and homogeneity, with their equations discussed below. Energy assesses the uniformity of the texture, while correlation examines the correlation between a pixel and its neighbor across the complete image. Dissimilarity quantifies the diversity in graylevel pairs, and homogeneity evaluates the proximity of the concentration of elements in the GLCM along its diagonal.Contrast gauges the local variations within the GLCM.

$$\text{Contrast} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-j)^2 P(i,j) \tag{1}$$

$$\text{Energy} = \sum_{i=o}^{N-1} \sum_{j=0}^{N-1} P(i,j)^2 \tag{2}$$

$$\text{Entropy} = -\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(i,j)\, log_2\, P(i,j) \tag{3}$$

$$\text{Dissimilarity} = \sum_{i=0}^{N-1} \sum_{j=o}^{N-1} |i-j|\, P(i,j) \tag{4}$$

$$\text{Correlation} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{(i-\mu_i)(j-\mu_j)\, P(i,j)}{\sigma_i \sigma_j} \tag{5}$$

Where μ_i, $\mu_j$are the mean and σ_i and σ_j are the standard deviations of each individual marginal distribution of i and $j.P(i,j)$ is the normalized value in the GLCM at position (i,j) , N is the number of Gray levels in the image.

3.3. Classification using Machine Learning Algorithms

The research employs a variety of well-established machine learning algorithms to categorize GLCM features extracted from images, showcasing their efficacy in diverse classification endeavors. In this research, five supervised ML classifiers were investigated: RF, SVM, KNN, DT, and NB. Chosen for their robust performance in image classification, these classifiers underwent optimization using the grid search method to attain the most favorable outcomes.

The research employs various ML algorithms for classification, with each algorithm being chosen for its distinctive strengths. RF constructs multiple decision trees and combines them for accurate and stable predictions. SVM with a linear kernel is utilized for its simplicity and effectiveness, especially in high-dimensional and binary classification tasks. KNN generates predictions based on the K in similar instances, making it straightforward to implement and effective for small datasets. DT models are selected for their ease of interpretation and visualization and their strong performance with smaller datasets.Lastly, Naive Bayes is a probabilistic model that leverages Bayes' theorem and assumes that features are independent of one another.

4. RESULTS AND DISCUSSION

In assessing the performance of a trained model, the statistical metrics including recall,precision, F1-score, accuracy, confusion matrix, and AUROC are employed. Precision evaluates the percentage of accurate predictionspositives, while recall represents the True Positive Rate (TPR).

4.1. Performance Metrics

The performance of the classifiers is assessed utilizing the following measures:

$$\text{Accuracy} = TP + TN/(TP + FP + TN + FN) \quad (6)$$
$$\text{Precision} = TP/(TP + FP) (7)$$
$$\text{Fallout} = FP/(FP + TN) (8)$$
$$\text{Recall} = TP/(TP + FN) (9)$$

True Positives are denoted by TP, True Negatives by TN, False Positives by FP, and False Negatives by FN.

4.2. Accuracy of Proposed Models

Accuracy measures the ratio of correctly identified instances to the overall number of instances presented in Table1.

Table 1. Comparison of Accuracy of different ML models.

ML algorithm Classifier	Training Accuracy	Testing Accuracy
RF	100%	99%
SVM	83%	58%
KNN	95%	81%
DT	100%	98%
NB	98%	100%

4.3. Performance Comparison

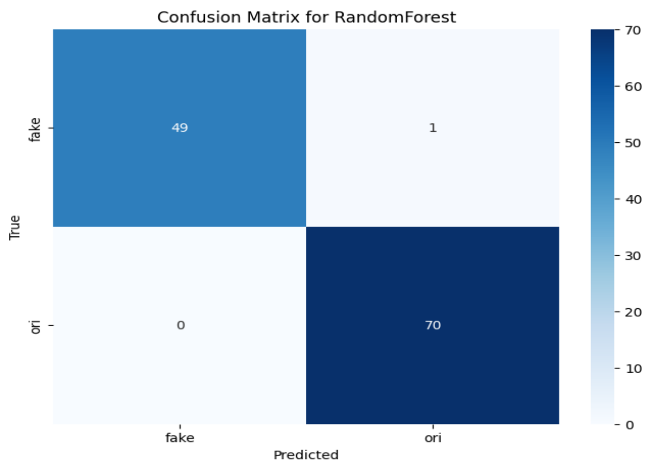
The RF and NB classifiers demonstrated the highest accuracy and AUC scores, indicating their superior performance in the classification of Caste certificate images. The robustness of the RF classifier stems from its ensemble nature and several decision trees. On the other hand, Naive Bayes is a classification method that relies on Bayes' theorem, assuming strong independence between features, and is computationally efficient and effective for certain data distributions. Decision Trees excel with GLCM features by capturing texture patterns and facilitating clear, rule-based decisions. In contrast, KNN and SVM exhibited relatively lower performance, encountering challenges when dealing with high-dimensional GLCM features and optimal hyperplane separation.

4.4. Confusion Matrices for Proposed Models

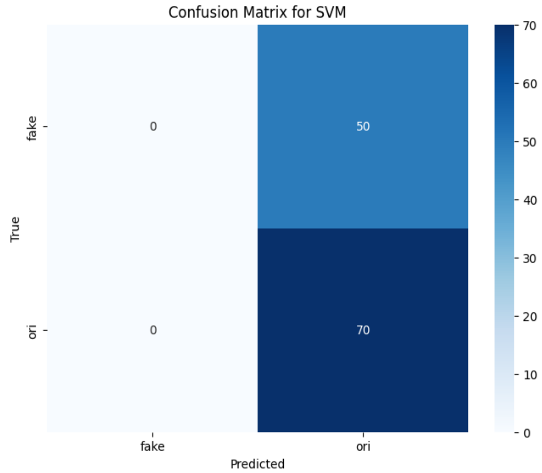
Confusion matrices provide comprehensive insights into classification performance by presenting the true-positive, false-positive, true-negative, and false-negativecounts for each classifier displayed in Table 2. Figure. 3 displays the confusion matrix for our proposed ML models, comparing the forecasted labels with the real labelsto evaluate the models' performance of test data.(a) RF (b) SVM (c) KNN (d) Decision tree (e) Naive Bayes

Table 2. Classificationperformance metrics for different ML models.

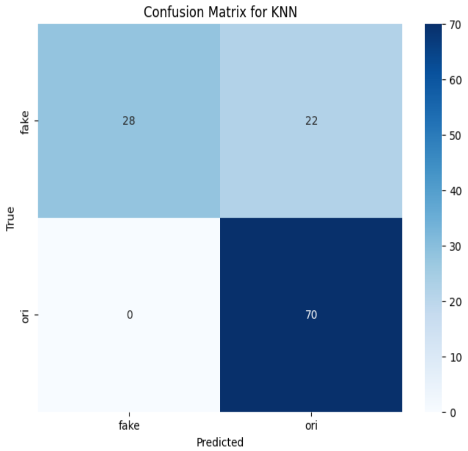
Classifier	TP	TN	FP	FN
RF	70	49	1	0
SVM	70	0	50	0
KNN	70	28	22	0
DT	70	48	2	0
NB	70	50	0	0



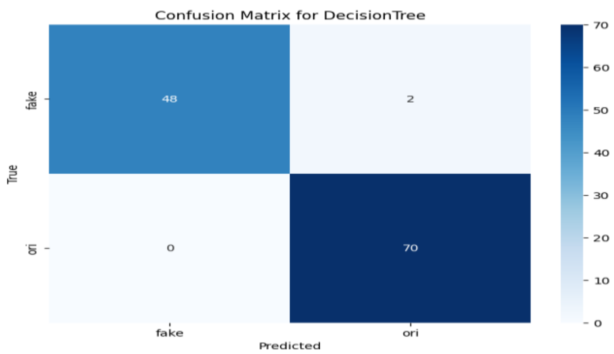
(a)


Confusion Matrix for SVM
True
fake
ori
Predicted
0
50
0
70
70
60
50
40
30
20
10
0

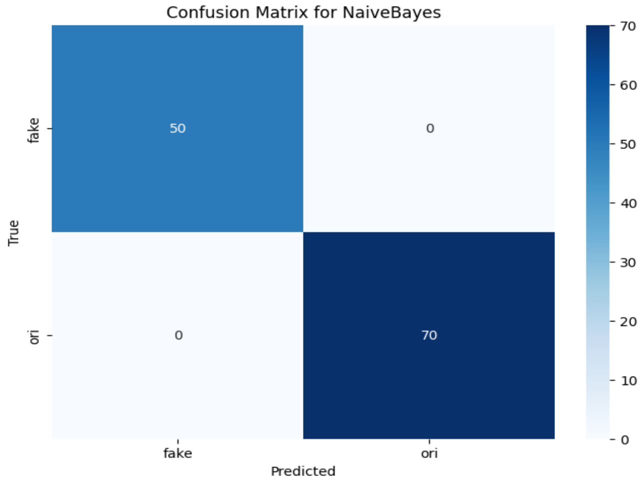
(b)


Confusion Matrix for KNN
True
fake
ori
Predicted
28
22
0
70
70
60
50
40
30
20
10
0

(c)


Confusion Matrix for DecisionTree
True
fake
ori
Predicted
48
2
0
70
70
60
50
40
30
20
10
0

(d)



(e)

4.5. Classification Report for Proposed ML Models

In machine learning classification, a classification report offers an overview of a model's performance, encompassing metrics that evaluate its accuracy in assigning data points to specific classes. The statistical analysis of proposed ML models ispresented in Table 3.

Table 3. Statistical results of the proposed ML models

ML Algorithms	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
RF	99	99	99	99
SVM	29	50	37	58
KNN	88	78	79	82
DT	99	98	98	98
NB	100	100	100	**100**

4.6. ROC AUC Scores for Proposed Models

The ReceiverOperating Characteristic furnishes a consolidated measure of performance across all classification thresholds. This metric is valuable for assessing the balance between the TPR and FPR. Table 4 presents elevated ROC AUC scores, indicating enhanced classifier performance and providing a comprehensive evaluation across all thresholds.Figure.4 displays the AUC-ROC curvesof ML models.

Table 4. AUC score of the proposed ML models

Classifier	AUC Score
RF	1.00
SVM	0.98
KNN	0.93
DT	0.98
NB	1.00

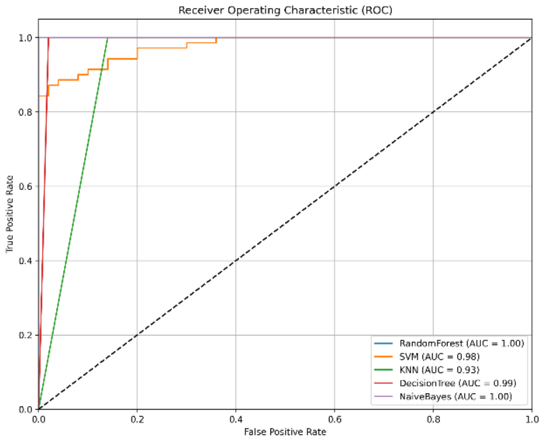


Figure4:Proposed ROC curve for the model

4.7. Hyperparameter Comparison for GLCM Features

In our experiments, the Naïve Bayes classifier delivered flawless accuracy (100%) with a var_smoothing parameter of 1.0, surpassing all other models. Random Forests also demonstrated strong performance, achieving 99% accuracy with a max_depth of 10 and n_estimators of 10. Decision Trees achieved 98% accuracy with a min_samples_split of 5. KNN and SVM were less effective, with KNN reaching 82% accuracy and SVM 58% accuracy using their optimal parameters. As shown in Table 2.

Table 5. Hyperparameter Comparison for GLCM FEATURES

Algorithm	**Best Parameters**	**Accuracy**
Random Forests	max_depth: 10	99
	n_estimators: 10	
KNN	n_neighbors: 3	82
	weights: distance	
Decision Trees	min_samples_split: 5	98
SVM	C: 0.1	58
	kernel: linear	
Naïve Bayes	nb__var_smoothing': 1.0	**100**
	'scaler__with_mean': True	

5. CONCLUSION

This study explores the classification of Indian Caste certificate images using GLCM features and a variety of ML algorithms. The key findings reveal that the RF, DT, and NB classifiers attained the highest accuracies of 99%, 98%, and 100%, respectively, with corresponding AUC scores of 0.99, 0.98, and 1. These results demonstrate their effectiveness in discriminating between original and counterfeit Caste certificate images. While the KNN and SVM classifiers exhibited relatively poorer performance, the GLCM features proved to be effective for texture analysis, providing valuable information for classification. The comparative analysis emphasized the superior performance of ensemble methods such as RF, DT, and NB. Despite promising results, the study's constraints involve having a relatively small dataset, which may impact generalization. Future work should validate the findings on larger datasets and explore additional features or deep learning techniques to further improve classification performance. This researchcontributes to document verification by using GLCM features and ML algorithms in classifying Caste certificate images, laying the groundwork for the development of automated systems to identify forged documents and ensure equitable benefit allocation.

ACKNOWLEDGMENT

The authors extend their appreciation to Directorate Nada Kacheri, Government of Karnataka, for providing Caste certificate datasets for this research work. The author Subramani C thanks the University Grants Commission (UGC) New Delhi, India, for providing a fellowship under the UGC JRF. scheme for research work. (UGC JRF Award letter No.: 210510302177).

REFERENCES

[1] Nath, Siddhartha Sankar, et al. "A survey of image classification methods and techniques." 2014 International conference on control, instrumentation, communication and computational technologies (ICCICCT). IEEE, 2014.
[2] Thakur, Nupur, and Deepa Maheshwari. "A review of image classification techniques." Int. Res. J. Eng. Technol 4.11 (2017): 1588-1591.
[3] Humeau-Heurtier, Anne. "Texture feature extraction methods: A survey." IEEE access 7 (2019): 8975-9000.
[4] Jardine, M. A., J. A. Miller, and Megan Becker. "Coupled X-ray computed tomography and grey level co-occurrence matrices as a method for quantification of mineralogy and texture in 3D." Computers & Geosciences 111 (2018): 105-117.
[5] Pantic, Igor, et al. "Gray level co-occurrence matrix algorithm as pattern recognition biosensor for oxidopamine-induced changes in lymphocyte chromatin architecture." Journal of theoretical biology 406 (2016): 124-128.
[6] De Siqueira, Fernando Roberti, William Robson Schwartz, and Helio Pedrini. "Multi-scale gray level co-occurrence matrices for texture description." Neurocomputing 120 (2013): 336-345.
[7] Ansari, Mohd Dilshad, Satya Prakash Ghrera, and Vipin Tyagi. "Pixel-based image forgery detection: A review." IETE journal of education 55.1 (2014): 40-46.
[8] Wang, Junwen, et al. "Fast and robust forensics for image region-duplication forgery." Acta Automatica Sinica 35.12 (2009): 1488-1495.
[9] Farid, Hany. "Image forgery detection." IEEE Signal processing magazine 26.2 (2009): 16-25.
[10] Hayat, Khizar, and Tanzeela Qazi. "Forgery detection in digital images via discrete wavelet and discrete cosine transforms." Computers & Electrical Engineering 62 (2017): 448-458.
[11] Kumari, Ch Usha, S. Jeevan Prasad, and G. Mounika. "Leaf disease detection: feature extraction with K-means clustering and classification with ANN." 2019 3rd international conference on computing methodologies and communication (ICCMC). IEEE, 2019.
[12] Mokhtar, Usama, et al. "SVM-based detection of tomato leaves diseases." Intelligent Systems' 2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS’2014, September 24-26, 2014, Warsaw, Poland, Volume 2: Tools, Architectures, Systems, Applications. Springer International Publishing, 2015.
[13] Pinto, Loyce Selwyn, et al. "Crop disease classification using texture analysis." 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2016.
[14] Singh, Aditi, and Harjeet Kaur. "Potato plant leaves disease detection and classification using machine learning methodologies." IOP Conference Series: Materials Science and Engineering. Vol. 1022. No. 1. IOP Publishing, 2021.
[15] Iqbal, Naveed, et al. "Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms." PeerJ Computer Science 7 (2021): e536.
[16] Pouyap, Mireille, et al. "Improved Bearing Fault Diagnosis by Feature Extraction Based on GLCM, Fusion of Selection Methods, and Multiclass-Naïve Bayes Classification." Journal of Signal and Information Processing 12.4 (2021): 71-85.

[17] Li, Dian, Cheng Wu, and Yiming Wang. "A novel iris texture extraction scheme for iris presentation attack detection." Journal of Image and Graphics 9.3 (2021): 95-102.

[18] K. Padmavathi and Maya V. Karki. “Texture Feature Extraction and Classification of Brain Neoplasm in MR Images using Machine Learning Techniques”, International Journal of Recent Technology and Engineering, Vol. 8, No. 5, pp. 1-9, 2020..

[19] Barburiceanu, Stefania, Romulus Terebes, and Serban Meza. "3D texture feature extraction and classification using GLCM and LBP-based descriptors." Applied Sciences 11.5 (2021): 2332.

[20] Kumar, Dharmender. "Feature extraction and selection of kidney ultrasound images using GLCM and PCA." Procedia Computer Science 167 (2020): 1722-1731.