National Research Journal of Information Technology & Information Science
Volume No: 12, Issue No: 2, Year: 2025 (July- December)
PP: 67-74
ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

COUNTERFEIT DETECTION IN EDUCATIONAL CREDENTIALS USING MACHINE LEARNING TECHNIQUES

Sarala M

Department of Computer Science and Applications, Bangalore University, Bangalore, India

Muralidhara B L

Department of Computer Science and Applications, Bangalore University, Bangalore, India

ABSTRACT

Forgery detection in educational credentials is a challenging task. Machine learning(ML) techniques are highly used in fraud detection and spam detection. Our primary objective is to detect counterfeit educational credentials using ML algorithms. In this research work, we conducted an empirical study as follows: (i) extracted the features of those credentials using the Gray Level Histogram Analysis (GLHA) metrics such as standard deviation, mean, skewness, kurtosis, and entropy, (ii) Trained the ML models with extracted features using different classification algorithms including Support Vector Machine(SVM), Logistic Regression, Decision Tree, Naïve Bayes, Random Forest, and K-Nearest Neighbours(KNN). (iii) Assessed the effectiveness of classification models using hyper-parameters. Random Forest got 99.38% accuracy, and outperformed well than other algorithms. SVM, Decision Tree, Logistic Regression, KNN, and Naïve Bayes got accuracies of 98.75%, 98.13%, 95.00%, 92.50%, and 90.00% respectively.

Keywords: machine learning, GLHA, counterfeit detection, supervised algorithms.

1. INTRODUCTION

Machine Learning (ML) techniques combine the three major domains such as computer science, mathematics, and statistics. It is widely used in medical image analysis in healthcare, student performance analysis in the education sector, machinery fault detection in industry areas, credit card fraud detection, and fake currency detection in the banking sector, and so on. The supervised approach is trained on labeled datasets, and the unsupervised algorithms are trained without labels using clustering algorithms. This study focused on the detection of counterfeit educational credentials using supervised ML algorithms. We extracted the features using GLHA methods to train the ML models.

The remaining sections are organized as follows: a comprehensive literature review is discussed in Section 2. Section 3 covers the methodology used for detection of counterfeit educational credentials. Section 4 discusses comparative analysis and findings of ML algorithms. Finally, Section 5 concludes the results and outlines the future work.

2. LITERATURE REVIEW

Machine Learning (ML) techniques have found wide applications, including classification, time series analysis, and fraud detection. Numerous studies have explored document classification using image datasets. This review investigated the studies related with healthcare and agriculture, time series analysis, and document classification using Optical Character Recognition(OCR). ML classifiers have been used to identify COVID-19 cases using features extracted from computed tomography (CT) images [1], distinguish non-donors from blood donors with cirrhosis, fibrosis, and hepatitis using the UCI-MLR dataset [2], and for binary classification tasks using metabolomics datasets [3]. In agriculture, soybean seed

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

varieties were classified using Random Forest, Naïve Bayes, SVM, and MLP [4]. Further, ML methods have been utilized to classify neutron stars and black holes, with performance evaluations reported [5].

A comparative analysis analyzing various ML algorithms across multiple tabular datasets demonstrated their effectiveness[6]. For network anomaly detection, experiments on the KDDCup99 dataset revealed that the Average One Dependence Estimator (AODE) outperformed other ML methods [7]. Weather prediction tasks, such as classifying rainy and non-rainy days in Maharashtra, have been addressed using SVM and ANN [8]. Similarly, Random Forest and ANN were applied to distinguish between normal and malicious network traffic [9]. In the medical domain, features derived from GLCM and GLRLM on lung CT scans were used to classify COVID and non-COVID cases [10].

In document classification, Stochastic Gradient Descent provided the top results in classification of biomedical documents [11]. SVM outperformed in categorizing research articles into business, social sciences, and scientific [12]. Patents were classified based on geographical location[13]. Additionally, ML algorithms categorized institutional documents [14], and banknotes [15]. Numerous studies extracted features using OCR, and used ML algorithms to recognize the documents [16-18].

3. MATERIALS AND METHODS

3.1 Materials

We scanned four hundred copies of the original choice-based credit system(CBCS) marks cards issued by Bangalore University. The documents were scanned using a FUJITSU scanner at 300 dpi (dots per inch) and stored as image files. In addition, two hundred fabricated credentials were generated using Adobe Photoshop. To ensure balance between the two classes, upsampling techniques were applied. For experimentation, the 320 samples for training and 80 samples for testing.

3.2 Methodology

We employed machine learning techniques to categorize educational credentials as either ‘genuine’ or ‘fake’. Our approach is illustrated in Figure. 1. The datasets were preprocessed by resizing the image dimension to 256 ^{x} 256. Further, we converted the RGB images to grayscale channels to extract the features using GLHA. Since it's not practical to use image datasets directly for training the ML models, we first preprocess the datasets to extract features using GLHA. These extracted features were then saved in a database as a ‘.CSV’ file. Ultimately, we developed ML models utilizing six different classifier algorithms to classify the credentials and assess their performance.

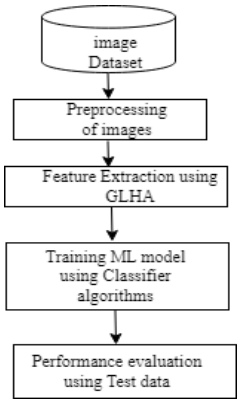


Figure 1. Proposed methodology

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

ML supervised algorithms:

Classifier algorithms play a crucial role in fitting machine learning models, utilizing labeled datasets for training. In this study, we implemented six classification algorithms: Support Vector Machine(SVM), Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Logistic Regression. SVM creates a hyperplane that effectively classifies data points by establishing boundaries, demonstrating efficiency in high-dimensional spaces. We selected a gamma (γ) value of 0.05 based on empirical research, and assessed SVM's performance by testing various kernel methods including 'linear', 'rbf', 'poly', 'sigmoid', and 'precomputed'. Naïve Bayes (NB) is widely employed in tasks like text and document classification, particularly in spam detection [19]. This research compared different Naïve Bayes variants. The Decision Tree algorithm constructs a tree structure using features and applies criteria such as 'entropy' and 'gini' for decision-making. Logistic Regression relies on the sigmoid function to estimate parameters, and its performance was evaluated using various 'solver' options, including 'lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky', 'sag', and 'saga', with 'newton-cg' yielding the best results. KNN, known for its 'lazy learning' approach. In this analysis, we fine-tuned the number of neighbors, ultimately choosing n=3. Lastly, an ensemble classification method Random Forest that operates multiple decision tree classifiers concurrently, also leveraging criteria like 'entropy' and 'gini' for improved accuracy.

3.3 Performance analysis

The performance of predictions are analyzed using statistical metrics as mentioned in Equations 1, 2, 3, and 4. In these equations, FP, TP, FN, and TN stand for false positive, true positive, false negative, and true negative, respectively.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

4. RESULTS AND DISCUSSIONS

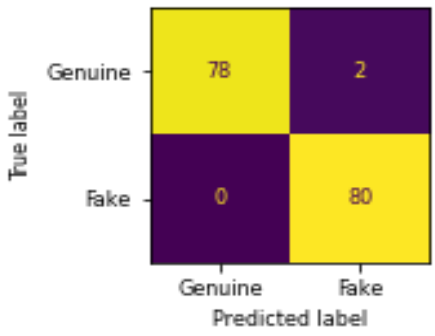
The effectiveness of machine learning algorithms after fine-tuning their hyperparameters is summarized in Table 1. The results for the six classifier algorithms are as follows: (i) Random Forest: This algorithm outperformed the others. (ii) SVM: The SVM with 'linear' and 'rbf' kernels showed superior performance compared to other kernels; however, the 'sigmoid' kernel did not fit the decision boundary as effectively, resulting in lower accuracy. (iii) Naïve Bayes: Gaussian Naïve Bayes outperformed the other variants, while Bernoulli Naïve Bayes is more efficient with discrete features but achieved lower accuracy overall. (iv) Decision Tree: Both the 'entropy' and 'gini' criteria demonstrated strong performance. (v) Logistic Regression: This algorithm delivered the best results, except for 'sag' and 'saga', which are more appropriate for multinomial logistic regression, thus yielding less accuracy for binary classification. (vi) KNN: The algorithm performed effectively with an odd number

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

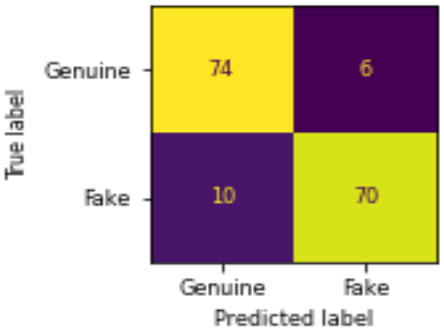
of neighbours, specifically using a value of 3 in this case. The confusion matrices for all ML algorithms are illustrated in Figures 2(a) to 2(f). The AUROC performance of ML algorithms is measured in percentage in Figure 3, and the accuracy metrics are illustrated in Figure 4.

Table. 1. Performance analysis

Sl. No	Classifiers	Method	Parameters	Accuracy
1.	SVM	Kernel	linear	98.8
			poly	97.5
			rbf	98.8
			sigmoid	50.0
2.	Naïve Bayes (NB)	Gaussian NB	var_smoothing =1e-09	90.0
		Multinomial NB		79.4
		Complement NB		79.4
		Bernoulli NB		50.0
3.	Decision Tree	Criterion	entropy	98.2
			gini	98.2
4.	Logistic Regression	Solver	newton-cg	95.0
			newton-	95.0
			lbfgs	95.0
			liblinear	95.0
			sag	80.0
			saga	79.4
5.	KNN	Number of neighbours	3	96.3
6.	Random Forest	Criterion	entropy	99.4
			gini	99.4

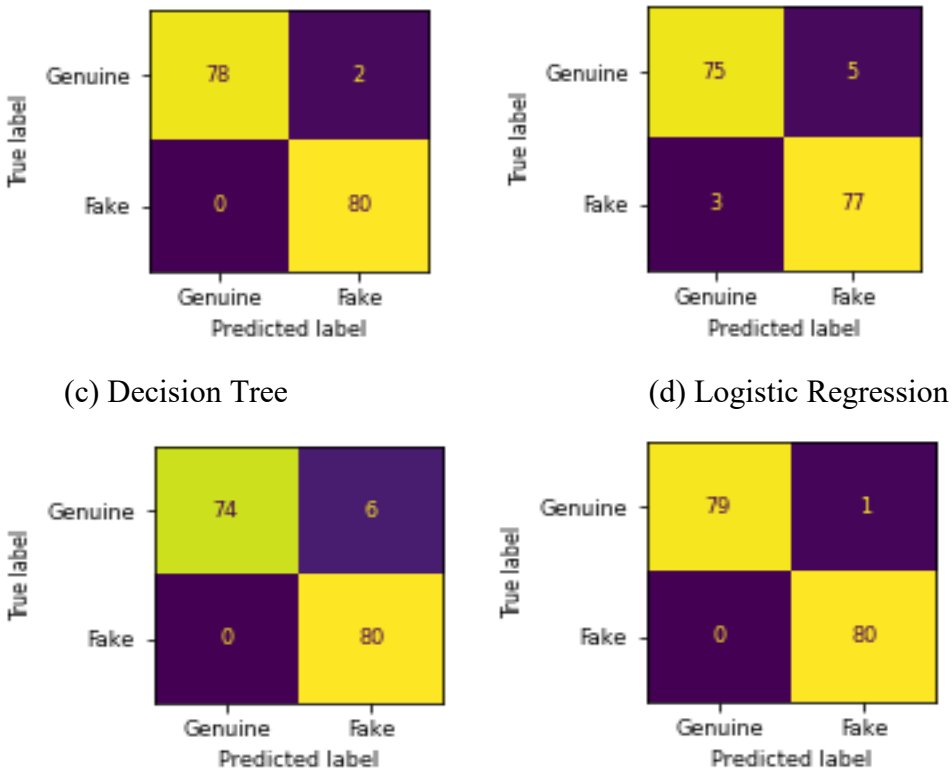


(a) SVM



(b) NaiveBayes

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in



(c) Decision Tree (d) Logistic Regression

(e) KNN (f) Random Forest

Figure 2. Confusion Matrix of ML algorithms

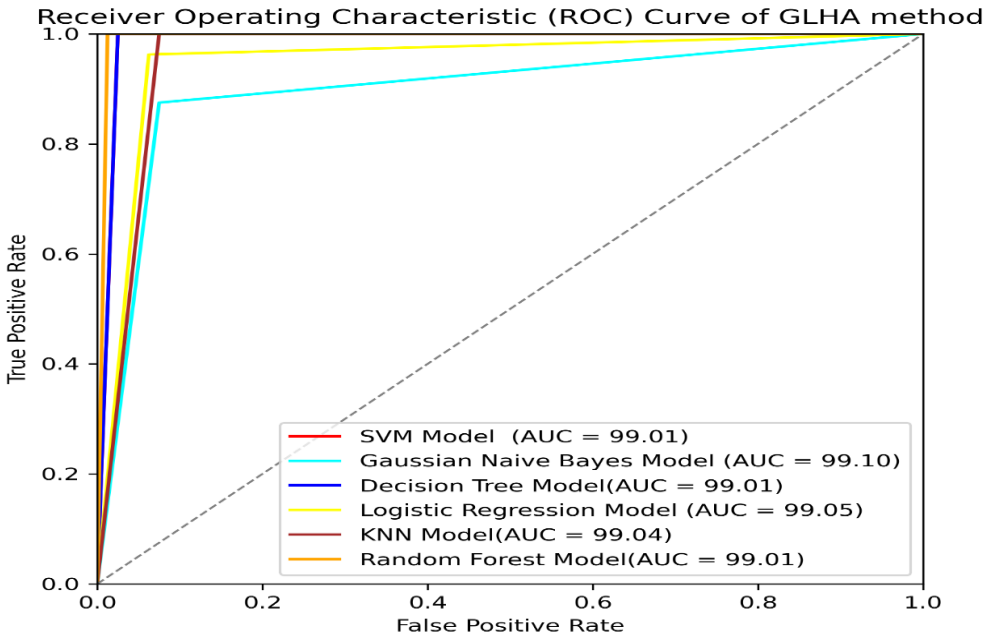


Figure 3. AUROC of ML algorithms

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

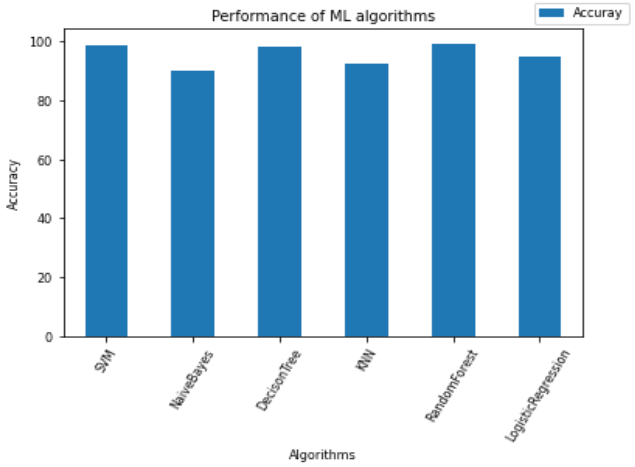


Figure 4. Performance analysis of ML algorithms

5. CONCLUSIONS

The GLHA method extracted the features from the educational credentials and then classified as either 'genuine' or 'fake' with the help of various classifiers. The effectiveness of ML algorithms utilizing GLHA was examined, along with their hyper-parameter settings, and the findings are discussed in Section 4. Among the algorithms, Random Forest and SVM with a 'linear', and 'rbf' kernel demonstrated superior performance, achieving accuracies of 99.38% and 98.75%, respectively. The Decision Tree, using both the 'entropy' and 'gini' criteria, reached an accuracy of 98.13%. Logistic Regression achieved an accuracy of 95.00%, while KNN recorded an accuracy of 92.5%. Naïve Bayes with the Gaussian Naïve Bayes variant attained an accuracy of 90.00%. However, a limitation of this study is that it does not predict fabricated credentials that may emergge in the future, as the machine learning models are trained on labeled datasets. Therefore, our future work will aim to detect counterfeit educational credentials using a semi-supervised approach with Convolutional Neural Networks (CNNs) to enable more generalized results.

ACKNOWLEDGEMENTS.

The authors would like to thank Bangalore University for providing the scanned datasets of marks card for this research work. The author Sarala M acknowledges the Department of Science and Technology(DST), Karnataka Science and Technology Promotion Society (KSTePS), Government of Karnataka, India for receiving the fellowship support (Award letter No.: DST/KSTePS/Ph.D. Fellowship/PHY-06: 2022 – 23/472) for research work.

REFERENCES

1. Godbin, A. Beena, S. Graceline Jasmine.: Screening of COVID-19 based on GLCM features from CT images using machine learning classifiers, SN Computer Science 4, no. 2, 133 (2022).
2. Fahad B Mostafa, Easin Hasan.: Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data, DOI:10.1101/2021.04.26.21256121, (2021).
3. Mendez, Kevin M., Stacey N. Reinke, David I. Broadhurst. : A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. Metabolomics 15:1-15. (2019).

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

4. Çetin, Necati.: Machine learning for varietal binary classification of soybean (Glycine max (L.) Merrill) seeds based on shape and size attributes. Food Analytical Methods 15, no. 8: 2260-2273 (2022).
5. de Beurs, Zoe L., N. Islam, G. Gopalan, S. D. Vrtilek.: A Comparative Study of Machine-learning Methods for X-Ray Binary Classification, The Astrophysical Journal 933, no. 1: 116 (2022).
6. Araujo Santos, Vitor Cirilo, Lucas Cardoso, Ronnie Alves. The quest for the reliability of machine learning models in binary classification on tabular data, Scientific Reports 13, no. 1: 18464 (2023).
7. Nawir, Mukrimah, Amiza Amir, Ong Bi Lynn, Naimah Yaakob, R.Badlishah Ahmad. : Performances of machine learning algorithms for binary classification of network anomaly detection system, In: Journal of Physics: Conference Series, vol. 1018, pp. 012015. IOP Publishing, (2018).
8. Hudnurkar, Shilpa, Neela Rayavarapu.: Binary classification of rainfall time-series using machine learning algorithms, International Journal of Electrical & Computer Engineering (2088-8708) 12, no. 2 (2022).
9. Kirichenko, Lyudmyla, Tamara Radivilova, Vitalii Bulakh.: Binary classification of fractal time series by machine learning methods, In: Lecture Notes in Computational Intelligence and Decision Making: Proceedings of the XV International Scientific Conference on Intellectual Systems of Decision Making and Problems of Computational Intelligence(ISDMCI'2019), Ukraine, May 21–25,15, pp. 701-711. Springer International Publishing, (2020).
10. Jhansi, B., M. Ramesh, A. Deepak, P. R. Karthikeyan.: Evaluating Textural Changes of Lung in CT Images using GLCM in Comparison with GLRLM. In: 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 857-862. IEEE, (2022).
11. Behera, B., Kumaravelan, G. (2020). Performance evaluation of machine learning algorithms in biomedical document classification. Performance Evaluation, 29(5): 5704-5716.
12. Chowdhury, S., Schoen, M.P. (2020). Research paper classification using supervised machine learning techniques. In 2020 Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, pp. 1-6. https://doi.org/10.1109/IETC47856.2020.9249211
13. Miric, M., Jia, N., Huang, K.G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. Strategic Management Journal, 44(2): 491-519. https://doi.org/10.1002/smj.3441
14. Krasnyanskiy, M.N., Obukhov, A.D., Solomatina, E.M. (2019, March). The algorithm of document classification of research and education institution using machine learning methods. In 2019 International Science and Technology Conference "EastConf", Vladivostok, Russia, pp. 1-6. https://doi.org/10.1109/EastConf.2019.8725319
15. Rashid, F., Gargaare, S.M., Aden, A.H., Abdi, A. (2022). Machine Learning Algorithms for Document Classification: Comparative Analysis. International Journal

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

of Advanced Computer Science and Applications, 13(4): 260-265. https://doi.org/10.14569/IJACSA.2022.0130430

16. Ling, X., Gao, M., Wang, D. (2020). Intelligent document processing based on RPA and machine learning. In 2020 Chinese Automation Congress (CAC), Shanghai, China, pp. 1349-1353. https://doi.org/10.1109/CAC51589.2020.9326579
17. Guha, A., Samanta, D. (2020). Real-time application of document classification based on machine learning. In Intelligent Computing Paradigm and Cutting-edge Technologies: Proceedings of the First International Conference on Innovative Computing and Cutting-edge Technologies (ICICCT 2019), Istanbul, Turkey, pp. 366-379. https://doi.org/10.1007/978-3-030-38501-9_37
18. Omurca, S.I., Ekinci, E., Sevim, S., Edinç, E.B., Eken, S., Sayar, A. (2023). A document image classification system fusing deep and machine learning models. Applied Intelligence, 53(12): 15295-15310. https://doi.org/10.1007/s10489-022-04306-5.
19. Sarker, Iqbal H.: Machine learning: Algorithms, real-world applications, and research directions, SN computer science 2, no. 3: 160 (2021).