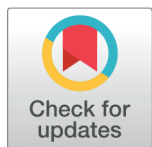


ORIGINAL ARTICLE



OPEN ACCESS

Received: 27/06/2026

Accepted: 11/09/2026

Published: 11/10/2026

Citation: Amin R, Bashir R (2026) Enhanced Script Identification for Multilingual Document Analysis: A Dataset and Machine Learning Approach for Kashmiri and English. Indian Journal of Science and Technology 19(35): 2486-2495. <https://doi.org/10.17485/IJST/v19i35.970>

* **Corresponding authors.**

rafia.amin@iust.ac.in

rumaan.bashir@islamicuniversity.edu.in

Funding: None

Competing Interests: None

Copyright: © 2026 Amin & Bashir. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](#))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Enhanced Script Identification for Multilingual Document Analysis: A Dataset and Machine Learning Approach for Kashmiri and English

Rafia Amin^{1*}, Rumaan Bashir^{1*}

¹ Department of Computer Science, Islamic University of Science and Technology, Pulwama, India

Abstract

Objectives: The main purpose of this work is to develop an automatic script identification system for multilingual documents by proposing a benchmark dataset of printed Kashmiri (Perso-Arabic) and English scripts and systematically evaluating HOG-based machine-learning models. The study aims to support multilingual document analysis and enhance the pre-processing step of Optical Character Recognition (OCR) systems. **Method:** This study has created a dataset of 4,423 printed text images from different publicly available academic sources, pre-processed the images by resizing and binarization, and extracted Histogram of Oriented Gradients (HOG) descriptors to represent script characteristics. Four machine learning classifiers – Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Random Forest (RF) were trained, and different evaluation metrics were used to evaluate their performance. **Findings:** The proposed HOG-SVM framework achieved an overall classification accuracy of **99.0%**, demonstrating superior performance compared with KNN, Naïve Bayes, and Random Forest classifiers, **and** also produced the fewest misclassifications, with **8 English-to-Kashmiri errors and only 4 Kashmiri-to-English errors**, demonstrating its effectiveness for Kashmiri-English script identification. **Novelty:** This work introduces a printed Kashmiri-English script benchmark dataset and a detailed baseline assessment of various machine learning classifiers in a low-resource scenario for script identification. The proposed optimized HOG-SVM-based framework is a reliable benchmark for further studies on Kashmiri script recognition and can serve as a foundation for developing a multilingual OCR system for low-resource languages.

Keywords: Kashmiri Language; HOG-ML; Perso-Arabic Script; Printed Text Dataset; Multilingual Document Analysis; Script Identification

1 Introduction

Script identification is a fundamental task in multilingual document analysis and OCR. It identifies the writing system of a document before performing subsequent tasks such as text recognition, machine translation, and information retrieval. Deep learning methods such as Convolutional Neural Networks (CNNs), lightweight architectures, and transformer-based models have shown significant improvements in the identification of widely used scripts due to their increasing importance in the field of multilingual document analysis, Optical Character Recognition (OCR) systems, and intelligent document processing systems^{(1), (2)}. But low-resource languages like Kashmiri are still largely unexplored. Kashmiri is the main language of the Kashmir Valley and is mostly written in the Perso-Arabic script. It shares many visual features with Arabic, Urdu, and Persian, but also has its own distinctive orthographic features like a large number of diacritical marks, context-dependent character shapes, and right-to-left writing as shown in Table 1. These features introduce fine-grained variations in character appearance and can make visual discrimination of Kashmiri from other visually similar scripts difficult^{(3), (4), (5)}. Such script-specific variations emphasize the need for models capable of learning discriminative visual patterns for reliable script identification. Furthermore, the scarcity of standardized Kashmiri datasets limits the development, fair comparison, and reproducibility of machine-learning and deep-learning models, thereby slowing progress in this important research area and posing significant challenges for automated script recognition^{(6), (7)}.

Table 1. Typical Characteristics of the Kashmiri Language

S. No.	Feature	Particulars
1	Writing/Reading Order	Right-Left
2	Diacritical Marks	Significant use of (Hamza, Tashdeed, Pesh, Zaer, Dots, Zabbar, etc.)
3	Complexity	More cursive than Devanagari & English.
4	Distinct Separation between Lines, Words & Letters	Very Rare

Early studies in the field mainly dealt with printed and scanned documents with handcrafted feature extraction techniques. Their applications have recently been expanded to handwritten documents, natural scene images, and multi-lingual script identification^{(8), (9)}. Multilingual script identification by a patch based approach with Non-Negative Matrix Factorization (NMF) has been suggested, in which document images are divided into smaller patches to get discriminative local features⁽¹⁰⁾. The method is able to reliably identify multiple scripts, by capturing script specific characteristics from local regions. These methods however, are significantly relying on hand-crafted feature representations and are not capable of capturing complex variations in the visual aspects of real-world document images. Line level script identification based on Chain Code Histogram (CCH) and Discrete Fourier Transform (DFT) features, and Support Vector Machine (SVM) classification, recognized an average of 95.14% of the line level scripts⁽¹¹⁾. These methods are good with regard to the accuracy of classification and have relatively low computational complexity, however, when there are variations in writing styles, fonts or the quality of the documents, the methods show limited performance. For natural scene images, two end-to-end solutions, Multi-Channel Mask (MCM) and Multi-Channel Segmentation (MCS), have been suggested. It has been demonstrated that MCS performs better than some of the existing methods, with a recall value of 54.34% on the ICDAR MLT 2017 dataset and 81.13% on MLe2e dataset⁽¹²⁾. Recent studies have explored deep learning based techniques for script identification in camera based document images, such as, ScriptNet, a two-stream CNN for this task⁽¹³⁾. The relatively low computational requirements of lightweight CNN architectures can make them suitable for resource-constrained applications, including mobile and real-time document processing.⁽¹⁴⁾

While deep learning has achieved significant progress in multilingual script identification using large annotated datasets⁽¹⁵⁾, low-resource languages like Kashmiri are still relatively under-explored. In particular, the efficacy of classical machine learning methods for printed Kashmiri script identification has not been studied systematically. To address this research gap, this work proposes a benchmark dataset for the printed Kashmiri-English script identification task and provides a detailed comparative evaluation of state-of-the-art machine learning classifiers using Histogram of Oriented Gradients (HOG) feature descriptors. The proposed framework provides a strong baseline for future work on Kashmiri script identification and development of multilingual OCR systems for low resource languages.

1.1 Research Gap

Most existing work on script identification has been on high-resource languages and relatively little attention has been given to low-resource languages. The absence of benchmark datasets and the scarcity of research in this area has limited the development, evaluation and comparison of different approaches for script identification in a fair manner. Moreover, the lack of studies that

systematically compare several traditional machine learning classifiers in a common experimental setting makes it difficult to provide reliable baseline results and reproducibility. This has led to limited progress in developing robust script identification systems for low-resource languages. Thus, benchmark datasets have to be created, and detailed baseline results have to be established for future research in script identification and multilingual document analysis.

1.2 Objectives of the study

To fill these research gaps, this work presents a new printed Kashmiri-English script dataset for script identification. The proposed dataset is tested with four traditional machine learning classifiers namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF) with Histogram of Oriented Gradients (HOG) feature descriptors. The main purpose of this study include:

- To develop a benchmark dataset for printed Kashmiri-English script identification;
- Systematically assess and compare performance of different machine learning classifiers under a common experimental setup;
- Identify the best classifier for Kashmiri-English script recognition that can serve as a reliable baseline for future research on Kashmiri Optical Character Recognition (OCR), multilingual document analysis, and other low-resource script identification tasks.

2 Methodology

2.1 Dataset Creation

A custom dataset containing 4,423 script images was built from publicly available academic resources. There are more than 2,000 samples in the dataset for each of the two script classes, English and Kashmiri as represented in Figure 1. The dataset is made up of printed text images as shown in Figure 2, to facilitate research on script identification in low-resource settings where large standardized datasets are often not available. The dataset was divided into training and testing sets to test the generalization ability of the models to unseen samples. The exact number of samples in each class is shown in Figure 1, which gives a clear picture of the distribution of the dataset. The organization of the dataset enables reliable evaluation of machine learning-based script identification methods.

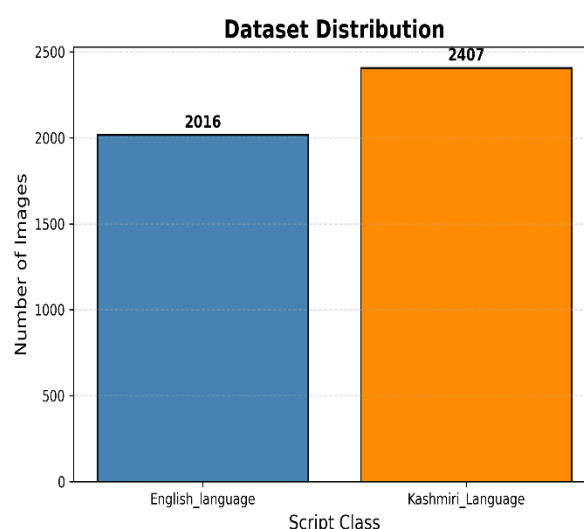


Fig 1. Dataset Distribution.

English gloss of Kashmiri Sample Image (2b): “Sheikh Ul Aalam (RA) was a great Islamic scholar and spiritual leader who devoted his life to worshipping Allah and spreading the teachings of Islam. He lived a simple and humble life and taught people about faith, honesty, kindness, and good character. He stood against injustice. Helped the poor and needy, and always tried to make things easier for people. His life remains an inspiration for us to live with simplicity, compassion, justice, and devotion to Allah”.

4.2. Introducing New Recipes. A single recipe can be cooked in numerous ways by combining the ingredients. In addition, the fact is that those ingredients can be cooked in some other ways which creates an area of endless possibilities to make cooking dishes. On the web, multiple recipes are available online and contain a huge dataset that allows the exploration of components in several cuisines from common man to professionals. The researchers can determine the differences as well as the similarities between the different cuisines [40]. For example, North American and Western European cuisines are totally based on the ingredients which contain the same flavor of compounds that East-Asian and Southern-European cuisines avoid.

In conclusion, the technologists can appraise which food components have a decent savor and mark a cuisine that is trendy in a few provinces. This basic understanding also permits artificial-intelligence-based algorithms to recommend the chefs with different types of ingredient combinations that will surely result in broadening the menu as well as the profits of a food industry.

حضرت شیخ العالیؒ اس کو کہہ چار تھے تکیس خدایہ بجز عبادت کر کے کچھ گل - واریتین
واریتین رو دو کہو، وکاس منفر سے یہ منفر الدنقانی پدیس عبادت منفر معقول - سو ہمہ منفر غیر تھہ
بہمیری تم سارے تھہیر یہ وچہ تھہ لائین اسلامک، پز رک، مبلہ ارک یہ سمہ ارک سق -
سارے و انس مند ارک سادگی سان - جہنک سمکن عین، پدہ پوشاک یہ لسن ہن اوس سادہ -
حضرت شیخ العالیؒ کو اسلامی ریہک قائم - اوندہن خاوش ہیر تیش ونان - ہن آکو
واریہ زاتھ یہ مرید - تھیر منفر چہ واریتین جائین جہنہن ورائین وچہدی زیارت یہ آستان -
جہنہن خاص یارن، خاوش یہ مریدین منفر چہ زلیہ سمہو ربابا نام الدین (ہمہ روتش: ہمہ
ژو، بابا زین الدین) (زآہ شاہ، عہد مقام)، بابا نصر الدین (ژرارہ شریف) یہ بابا تاج
الدین (وڑتیل)۔

شیخ العالم آسکر زائمانی، پیدائشی مغربی مغلطہ خٹوٹوں بندہ کی کافی رول۔ جہند من
شاگردن مغربی آسبہ شام بی بی، دہشت دہ، بہت دہ چہرہ ناماء از غوثیہ شامیہ بی بی از غوثیہ گور
نامہ شہزادہ۔

شیخ الاسلام اُسی فلسفے سے جس خلافِ - جم اُسی عربین ہندی قوم خوار - تنہا لڑتے تھے
سارے خلافِ یمن عربین وصلِ نثر تھے لہٰذا اُسی - تھے اُس مائیں ہندی قومِ نعت - پر تھے
کائنات اُسی تھے روزگار یا کمال کمال تائید کران - لائیں راحت و اتقان اُسی عبادت

(a) English sample

(b) Kashmiri sample

Fig 2. Sample Images of dataset.

2.2 Proposed Methodology

2.2.1. Data Pre-processing

The dataset undergoes a set of pre-processing methods to prepare it for script identification and to ensure consistency, quality, and usefulness of the text images. These steps include:

- **Transformation to Image Format**

The PDFs from the sources were converted to high resolution image files. This step set the stage for later processing to accurately extract text without compromising the visual integrity of the source materials.

- **Text Region Extraction**

The document images were pre-processed to extract text regions and remove non-text regions. This step is important to reduce irrelevant information and improve the effectiveness of the subsequent script identification process.

- **Image Standardization**

The extracted text snippets were normalized to fulfill the input requirements of the script identification model. To maintain uniformity in the size of all images, all images were resized to a fixed resolution of 128x128 pixels. Other procedures were used to make the text more understandable such as:

Noise Reduction: – Removal of visual distractions and artifacts to improve overall quality. Recent studies highlight the importance of pre-processing for artifact removal, such as frequency-domain filtering in deraining task⁽¹⁶⁾.

Contrast Enhancement: Adjusting the brightness and contrast of an image can help to improve the visibility of text regions, especially in low quality or degraded document images.

2.2.2. Feature Extraction

The discriminative features of the pre-processed document images were obtained using the Histogram of Oriented Gradients (HOG). HOG is a hand-crafted feature descriptor which shows the distribution of local intensity gradients and edge directions in an image. It is particularly suitable for printed text material with structural properties like stroke patterns, character shapes and edge orientations. HOG yields an effective and compact representation that is able to capture the differences well and hence helps in accurate script identification⁽¹⁷⁾.

Before feature extraction, each document image was converted to grayscale, and resized to 128×128 pixels. The HOG descriptors were calculated using 9 orientation bins, cell size of 8×8 pixels, block size of 2×2 cells, and L2-Hys block normalization. Then, the extracted HOG feature vectors were standardized using StandardScaler to normalize the feature values

before classification. The normalized feature vectors thus obtained were used as input for all machine learning classifiers tested in this study.

2.2.3. Machine learning Classifiers

Machine learning classifiers were selected for comparative evaluation of HOG features using different classification approaches. The selected classifiers include: Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN) and Naïve Bayes (NB), which represent margin-based, ensemble, distance-based and probabilistic learning approaches, respectively. This selection allows the discriminative power of the HOG feature representation to be evaluated across classifiers with different learning mechanisms, thus providing a more comprehensive basis for comparing their effectiveness in printed script identification. Pre-processed document images were used to extract Histogram of Oriented Gradients (HOG) feature vectors, which were used to train each classifier. To ensure a consistent evaluation process for all the classifiers, the dataset was split into 80% training and 20% testing subsets. Hyperparameter optimization for all classifiers was done using GridSearchCV with five-fold cross-validation, and the best parameter combination from the grid search was used to train each final classifier.

- **Support Vector Machine (SVM)** is a supervised machine learning algorithm that is used for classification problems and can form a best decision boundary that maximizes the separation between classes. This property makes SVM particularly appropriate for a high dimensional feature space, so that it can achieve good generalization performance⁽¹⁸⁾. Histogram of Oriented Gradients (HOG) feature vectors were used in training the SVM classifier in this study. The HOG descriptors represent the structure of the script, and the SVM model is trained to discriminate between the scripts in order to achieve accurate identification. The trained classifier was then used to classify the input document images to their respective script categories.
- **K-Nearest Neighbors (KNN)** K-Nearest Neighbours (KNN) is a supervised non-parametric machine learning algorithm which classifies an input sample based on the majority class of its k nearest neighbours in the feature space. The simplicity and effectiveness of the classifier depends on the distance metric used to determine the similarity between the samples, which is the reason for the simplicity and effectiveness of this approach⁽¹⁹⁾. Pre-processed Document Images were used to extract the HOG feature vectors to train the KNN (k-Nearest Neighbor) classifier. In the HOG descriptors, the structural features of the scripts are encoded and when an image is presented to the KNN model, it is classified based on the feature representation of the image and the feature representation of the nearest training sample.
- **Random Forest (RF)** is a powerful ensemble learning algorithm which constructs multiple decision trees in training and classifies by voting the maximum of individual trees⁽²⁰⁾. The Random Forest takes the output from many trees to get a more accurate and robust classification than a single decision tree and less likely to over-fit. The Histogram of Oriented Gradients (HOG) feature vectors from the pre-processed document images were used to train the Random Forest Classifier. The HOG descriptors are the structural properties of the scripts, which can be used to effectively learn the discriminative patterns for script identification. The input document images were then classified to the script classes using the trained Random Forest model.
- **Naïve Bayes (NB)** is a kind of supervised machine learning algorithm that relies on the principle of conditional independence of features given the class label, and on the Bayes' theorem. Although this assumption is a simplification, Naïve Bayes has been demonstrated to have good performance in many text classification and pattern recognition problems because of its computational efficiency and its high-dimensional feature space tolerance⁽²¹⁾. In this work we used the Histogram of Oriented Gradients (HOG) feature vectors of the pre-processed document images for training the Naïve Bayes classifier. The extracted HOG descriptors represent discriminative information about the structure and the classifier can estimate the posterior probability of each script class and classify the input image to the class with highest posterior probability.

2.2.4. Implementation Details

The classical machine-learning models were implemented in Python using the Scikit-learn library. The images of the input document were converted to grayscale and were resized to (128×128) pixels. The HOG features were extracted with 9 orientations, (8 x 8) pixels per cell, (2 x 2) cells per block, and L2-Hys block normalization. The extracted HOG feature vectors were standardized using StandardScaler, and the class labels were encoded using LabelEncoder. The dataset was divided into training and held-out test subsets using an 80:20 stratified split. Then, the classifiers, namely SVM, Naïve Bayes, Random Forest, and KNN, were developed with their respective Scikit-learn implementation. For KNN, dimensionality reduction based on PCA was also added in a sequential pipeline with feature scaling. For each model, the best hyperparameter configuration was determined by GridSearchCV, with five-fold cross-validation and accuracy as the scoring criterion. The search spaces and selected configurations are summarized in Table 2. After optimization, the selected models were trained on the training subset

and evaluated on the held-out test subset with commonly used metrics for classification, including accuracy, precision, recall, F1-score, confusion matrix. These evaluation measures give the complete picture on the performance of the classifiers and can also facilitate a fair comparison among the classifiers evaluated.

2.2.5. Hyper-Parameter Configuration

The hyper-parameters of the machine learning classifiers were optimised via GridSearchCV. The search space and the selected configuration are shown in Table 2.

Table 2. Hyper-Parameter configuration of different Models

Model	Hyper-Parameters	Best Configuration
SVM	$C=\{0.1, 1, 10, 100\}$; kernel = {linear, RBF}; gamma = {scale, auto}	Linear kernel, $C=0.1$
Naïve Bayes	var_smoothing optimized using GridSearchCV with five-fold cross-validation	var_smoothing = 1×10^{-12}
Random Forest	Splitting criterion; maximum tree depth; minimum samples to split a node; minimum samples in a leaf node; number of decision trees	Gini criterion; 200 trees; max_depth = 20; min_samples_split = 2; min_samples_leaf = 1
KNN	Number of neighbors; distance metric; weighting scheme; number of PCA components; GridSearchCV with five-fold cross-validation.	9 neighbors; Euclidean distance; distance-based weighting; 150 PCA components

3 Results and Discussion

The document images were pre-processed, and Histogram of Oriented Gradients (HOG) feature descriptors were extracted to train and evaluate various machine learning classifiers for Kashmiri-English script identification. The Support Vector Machine (SVM) obtained the best overall performance among the classifiers considered. The optimized SVM model achieved a cross-validation accuracy of $98.82\% \pm 0.36\%$ and a test accuracy of 99.0%. This indicates a good generalization across different subsets of the dataset. The classifier achieved precision values of 0.99 and 0.98, recall values of 0.98 and 0.99 for English and Kashmiri, respectively, and F1-scores of 0.99 for both English and Kashmiri scripts (shown in Table 3), indicating the classifier's balanced and reliable classification performance. The robustness of the classifier is also verified by the confusion matrix shown in Figure 3 (a), where only 8 images of English script are misclassified as Kashmiri and 4 images of Kashmiri script are misclassified as English. The better performance of SVM can be attributed to its ability to separate high-dimensional HOG feature vectors and hence it is best suited for printed script identification.

The performance of the Naïve Bayes (NB) classifier was also evaluated using the extracted HOG feature descriptors. The optimized NB model obtained a cross-validation accuracy of $97.33\% \pm 0.67\%$ and a test accuracy of 97.0%. The Precision, Recall and F1-score of English script include: 0.99, 0.95, 0.97 and for Kashmiri script, the values are: 0.96, 0.99, 0.98 represented in Table 3. The confusion matrix in Figure 3 (b) indicates that 20 English script images are misclassified as Kashmiri and 3 Kashmiri script images are misclassified as English. Naïve Bayes classifier has good performance for printed script identification. These results show that the performance of Naïve Bayes is slightly lower than that of the SVM classifier. This is due to the conditional independence assumption of Naïve Bayes, which limits its ability to model the complex relationships among HOG features.

The Random Forest (RF) classifier was also evaluated using the extracted HOG feature descriptors. Using these settings, the optimized RF classifier obtained a cross-validation accuracy of $97.22\% \pm 0.69\%$ and a test accuracy of 98.00%. The classifier achieved precision of 0.99 and 0.96, recall of 0.95 and 0.99 and F1-scores of 0.97 and 0.98 for English and Kashmiri scripts (shown in Table 3), respectively. The confusion matrix in Figure 3 (c) shows that 19 English script images are misclassified as Kashmiri and only 3 Kashmiri script images are misclassified as English. The close agreement between the cross-validation and test accuracies shows that the model generalised well to unseen data. The Random Forest classifier was good in terms of classification, but its overall predictive performance was a little bit lower than that of the SVM classifier.

Another classifier tested in this study was the K-Nearest Neighbors (KNN) classifier, the baseline model achieved 58% accuracy. To improve performance, the dimensionality of the HOG feature vectors was reduced by applying Principal Component Analysis (PCA) to remove redundant and less informative components. With these settings the optimized model scored a cross-validation accuracy of $97.58\% \pm 0.40\%$ and a test accuracy of 98 %. The model achieved precision values of 0.99 and 0.97, recall values of 0.96 and 0.99, and F1-scores of 0.97 and 0.98 for the English and Kashmiri scripts (shown in Table 3) respectively. The significant improvement over the baseline demonstrated that PCA was able to minimize the effect of the high-dimensional HOG feature space on the distance-based KNN classifier, resulting in a more reliable and accurate script identification.

Table 3. Classification report of the implemented models

Model	Class	Precision	Recall	F1-Score	Accuracy
SVM	English	0.99	0.98	0.99	0.99
	Kashmiri	0.98	0.99	0.99	
KNN	English	0.99	0.96	0.97	0.98
	Kashmiri	0.97	0.99	0.98	
Naïve Bayes	English	0.99	0.95	0.97	0.97
	Kashmiri	0.96	0.99	0.98	
Random Forest	English	0.99	0.95	0.97	0.98
	Kashmiri	0.96	0.99	0.98	

The comparative evaluation of the machine learning classifiers shows that the Support Vector Machine (SVM) performed best in the Kashmiri-English script identification. The optimized SVM model achieved the best test accuracy of 99.0% and cross-validation accuracy of $98.82\% \pm 0.36\%$ which shows excellent generalization and the least misclassification errors. The Random Forest and the PCA-enhanced K-Nearest Neighbors (KNN) classifiers also performed competitively, both of them reaching a test accuracy of 98.0%. Naïve Bayes had the lowest performance among the evaluated classifiers, with a test accuracy of 97.0%. The results in general show that the HOG feature descriptors perform very well for printed Kashmiri-English script identification and SVM is found to be the most robust and accurate classifier.

3.1 Error Analysis

The confusion matrices in Figure 3 give us more information about the misclassification patterns of the evaluated classifiers. Overall, the models predicted most of the samples correctly in both English and Kashmiri and are able to discriminate between both scripts. The Naïve Bayes classifier misclassified 20 English samples as Kashmiri and only 3 Kashmiri samples as English. Random Forest misclassified 19 English samples as Kashmiri and 3 Kashmiri samples as English. KNN had 15 English-to-Kashmiri and 5 Kashmiri-to-English errors. In contrast, SVM produced the fewest English-to-Kashmiri errors, with 8 misclassifications, and 4 Kashmiri-to-English errors. The small number of errors across all models shows that HOG descriptors can learn discriminative structural information from printed script images.

Nevertheless, the residual misclassifications are due to the variations in character shapes, stroke structures, spacing, font properties, and other visual characteristics of the document images. Such differences can lead to similar local gradient and edge patterns in HOG representations, which might decrease the discrimination between the two scripts for some samples. In particular, there is the possibility that differences in character appearance and stroke arrangement may cause individual Kashmiri samples to show visual patterns that overlap with those learned from English samples, and vice versa. This overlap can lead to the cross-script misclassifications shown. For most classifiers, the higher number of English-to-Kashmiri errors compared to Kashmiri-to-English errors also suggests that some English examples have structural patterns that are less distinctive in the HOG feature space. However, all classifiers were able to separate the two scripts with high accuracy, and SVM yielded the most consistent separation among all techniques explored. The effectiveness of the proposed approach is further evaluated by comparing the proposed approach with the published studies in the following subsection.

3.2 Comparative Analysis with Published Studies

To further assess the effectiveness of the proposed approach, the performance of the proposed HOG-SVM framework was compared with representative script identification methods reported in the literature, as summarized in Table 4. Most existing studies have focused on high-resource languages and different types of datasets, whereas the present work focuses specifically on printed Kashmiri-English script identification using a dedicated low-resource dataset. Since the published studies were conducted using different datasets, script classes, pre-processing procedures, image types, and experimental protocols, their reported performance values are not considered for direct quantitative comparisons. Instead, they are presented to provide contextual insight into the competitiveness of the proposed framework.

Within the present experimental setting, SVM, KNN, Naïve Bayes, and Random Forest were independently implemented and evaluated on the curated dataset described in this study. Evaluating all classifiers under the same dataset and experimental conditions enabled a fair and consistent comparison of their performance. Among the evaluated classifiers, HOG-SVM achieved the highest performance, with an overall test accuracy of 99.0%, and was therefore selected as the proposed framework. The proposed HOG-SVM framework combines handcrafted Histogram of Oriented Gradients (HOG) features with an SVM

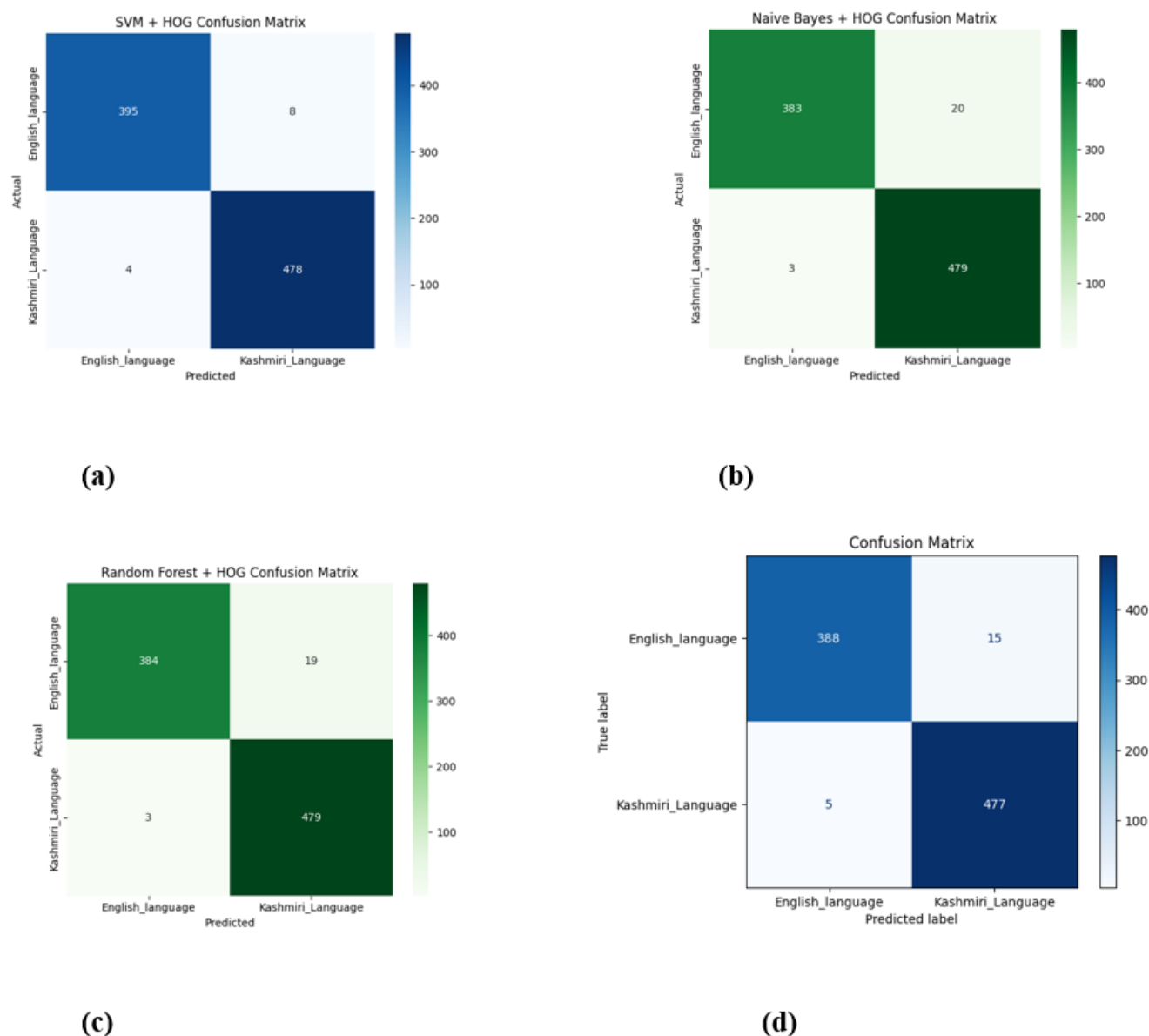


Fig 3. Confusion Matrices of different Models

classifier, demonstrating that carefully designed handcrafted features coupled with a traditional machine-learning classifier can achieve very high script-identification accuracy without the additional architectural complexity and computational requirements associated with deep-learning models.

For comparison with published studies, Mukherjee and Roy⁽⁶⁾ reported an accuracy of 99.2% using a CNN-BLSTM-CTC framework for simultaneous script identification and OCR. Singh et al.⁽¹¹⁾ reported an accuracy of 95.14% using SVM for line-level script identification, while Mahajan and Rani⁽²²⁾ reported an average F1-score of 94% using a CNN-based approach for script identification from scene images. The proposed HOG-SVM model achieved an accuracy of 99.0% and is therefore competitive with the reported results, particularly considering that it is obtained for a low-resource Kashmiri-English setting using handcrafted features and a conventional machine-learning classifier. However, owing to the differences in datasets, script classes, pre-processing procedures, and evaluation protocols, these results should be interpreted as contextual comparisons rather than direct quantitative comparisons. Overall, the findings demonstrate the effectiveness of HOG-SVM for

printed Kashmiri–English script identification and highlight its ability to achieve high accuracy with a comparatively simple classification framework.

Table 4. Comparison of the proposed HOG–SVM framework with published script identification methods.

Reference	Dataset	Method	Reported Performance	Remarks
Mukherjee and Roy ⁽⁶⁾	ISIDDI, RETAS, and a custom Hindi–Tamil dataset	CNN + BLSTM + CTC	Script Identification Accuracy = 99.2%	Simultaneous Script Identification and OCR
Singh et al. ⁽¹¹⁾	Indic Scripts	SVM	Accuracy = 95.14%	Script Identification at line-level
Mahajan and Rani ⁽²²⁾	NITJDATASET	CNN	Average F1-Score = 94%	Script Identification from Scene Images
Proposed Model	English-Kashmiri	HOG-SVM	Accuracy = 99%	Script Identification of Printed Document Images.

4 Conclusion

This study introduces a framework for identification of printed Kashmiri–English script using Histogram of Oriented Gradient (HOG) feature descriptors. In this study, the goal was to build a robust script identification system for a low resource language by creating a printed Kashmiri–English dataset and by evaluating several machine learning classifiers in a systematic way under a common experimental framework. Among the classifiers evaluated, the proposed HOG-SVM model yielded the best results with a cross-validation accuracy of $98.82\% \pm 0.36\%$, test accuracy of 99.0%, and precision and recall of 0.99 and 0.98 for English and 0.98 and 0.99 for Kashmiri, respectively. These results show that HOG descriptors are able to capture the distinctive structural features of the two scripts using a traditional machine learning approach.

The primary contribution of this work is a printed Kashmiri–English script dataset and an extensive comparative evaluation of classical machine learning classifiers using different learning mechanisms for script identification in the context of a low resource language. Results show that handcrafted HOG features combined with the SVM classifier offer a strong and computationally efficient baseline with competitive performance, without the complexity of deep learning models. This work provides a strong benchmark for future research on Kashmiri script identification and multilingual document analysis.

Future work can extend the proposed framework for multilingual script identification by incorporating more Indic and Perso-Arabic scripts and evaluating the performance on handwritten documents, degraded historical manuscripts, and complex document layouts. Furthermore, advanced architectures, including transformer-based approaches such as Vision Transformers (ViTs) and OCR-based models such as TrOCR, PaddleOCR, and CRNN, could be explored to enhance the recognition performance further. Since the current dataset contains approximately 300 more Kashmiri images than English images, future versions of the proposed Kashmiri–English dataset will also aim for a more balanced class distribution. A balanced and publicly available benchmark dataset will help the development, fair comparison, and evaluation of future script identification methods. These developments will ease the development of reliable OCR systems, digital document preservation, and document analysis for low-resource languages.

5 Acknowledgements

No funding was received for this article.

Data Availability : The dataset used in this study is not publicly available because it forms part of ongoing research.

Author Contribution: **Rafia Amin:** Conceptualization, Data collection and pre-processing, Model development, Experiment execution, Drafting the original manuscript. **Dr. Rumaan Bashir:** Supervision, Guidance on methodology and experimental design, Critical review and revision of the manuscript.

References

- 1) Moitra M, Saha SK. A review on handwritten text segmentation in Indian languages. *International Journal of Machine Learning and Cybernetics*. 2025 Jun;16(5):3269–3295. Available from: <https://link.springer.com/article/10.1007/s13042-024-02448-1>.
- 2) Ferrer MA, Das A, Diaz M, Morales A, Carmona-Duarte C, Pal U. MDIW-13: a new multi-lingual and multi-script database and benchmark for script identification. *Cognitive Computation*. 2024 Jan;16(1):131–157. <https://doi.org/10.1007/s12559-023-10193-w>.
- 3) Kashyap AK, Upadhyay M, Panwar VS, Chandrakar V. Integrated framework utilizing scene text detection and recognition techniques for enhancing point of interest extraction from name boards in all Indic languages. *Scientific Reports*. 2026 Mar.
- 4) Fayaz SA, Khaja MM, Bhat AS, Mansoor D, Thapa A, Zaman M. A baseline optical character recognition framework for printed Kashmiri Nastaliq text using deep learning. *Acadlore Transactions on AI and Machine Learning*. 2026;5(2):119–137. <https://doi.org/10.56578/ataiml050203>.

- 5) Bashir R, Quadri SMK. Density based script identification of a multilingual document image. *Int J Image Graph Signal Process*. 2015;2:8–14. <https://doi.org/10.5815/ijigsp.2015.02.02>.
- 6) Mukherjee J, Roy U. A low resource multi-lingual simultaneous script identification and text recognition model. *SN Computer Science*. 2024 Jul;5(6):740. <https://doi.org/10.1007/s42979-024-03107-6>.
- 7) Deyar DU, Ramani A, Gupta D, Nair PC, Venugopalan M. Dataset creation and benchmarking for Kashmiri news snippet classification using fine-tuned transformer and LLM models in a low resource setting. *Scientific Reports*. 2025 Nov;15(1):40828. <https://doi.org/10.1038/s41598-025-24451-4>.
- 8) Verma M, Sood N, Roy PP, Raman B. Script identification in natural scene images: a dataset and texture-feature based performance evaluation. Springer Singapore. 2016. https://doi.org/10.1007/978-981-10-2107-7_28.
- 9) Naosekham V, Sahu N. A hybrid scene text script identification network for regional Indian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 2024 Aug;23(8):1–26. <https://doi.org/10.1145/3649439>.
- 10) Arabnejad E, Moghaddam RF, Cheriet M. PSI: Patch-based script identification using non-negative matrix factorization. *Pattern Recognition*. 2017;67:328–339.
- 11) Singh PK, Chatterjee I, Sarkar R, Nasipuri M. Handwritten Script Identification from Text Lines. *Communication and Power Engineering*. 2017 Feb;p. 1. Available from: <https://arxiv.org/pdf/2009.07433>.
- 12) Khalil A, Jarrah M, Al-Ayyoub M, Jararweh Y. Text detection and script identification in natural scene images using deep learning. *Computers & Electrical Engineering*. 2021 May;91:107043. <https://doi.org/10.1016/j.compeleceng.2021.107043>.
- 13) Deng M, Ma H, Liu L, Qiu T, Lu Y, Suen CY. Scriptnet: a two stream CNN for script identification in camera-based document images. Springer Nature Singapore. 2022. Available from: https://link.springer.com/chapter/10.1007/978-981-99-1645-0_2.
- 14) Dargan S, Kumar M, Ayyagari MR, Kumar G. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering*. 2020 Sep;27(4):1071–1092. <https://doi.org/10.1007/s11831-019-09344-w>.
- 15) Yang Y, Eli E, Aysa A, Ubul K. Script identification in multilingual environment: a survey in recent years. *Artificial Intelligence Review*. 2025 Jul;58(10):294.
- 16) Shao M, Bao Z, Liu W, Qiao Y, Wan Y. Frequency domain-enhanced transformer for single image deraining. *The Visual Computer*. 2024 Oct;40(10):6723–6738. <https://doi.org/10.1007/s00371-023-03252-8>.
- 17) Dalal N, Triggs B. Histograms of oriented gradients for human detection;vol. 1. IEEE. 2005.
- 18) Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20(3):273–297. Available from: <https://link.springer.com/article/10.1007/BF00994018>.
- 19) Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*. 1992;46(3):175–185.
- 20) Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32. Available from: <https://link.springer.com/article/10.1023/A:1010933404324>.
- 21) Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*. 1997;29(2):103–130.
- 22) Mahajan S, Rani R. Word level script identification using convolutional neural network enhancement for scenic images. *Transactions on Asian and Low-Resource Language Information Processing*. 2022 Mar;21(4):1–29. <https://doi.org/10.1145/3506699>.