Machine Learning-based Signature Verification using OCR and Line Sweep Technique

Rajeev Gowda R., Hemanth Kumar, Karthik M, Nandan G

Dept. of Electronics and Communication Engineering ATME College of Engineering, Mysuru, India

DOI: https://doi.org/10.51244/IJRSI.2026.1306000429

Received: 25 June 2026; Accepted: 30 June 2026; Published: 16 July 2026

ABSTRACT

Handwritten signatures remain a widely used biometric for identity verification in banking, legal and governmental domains. Manual verification is slow and error-prone; automated verification is necessary to reduce fraud and speed processing. This paper presents a hybrid machine-learning based offline signature verification system that integrates Optical Character Recognition (OCR) for signature localization, Connected Components analysis and a Line Sweep geometric feature extraction technique, combined with supervised classifiers (Support Vector Machine and Convolutional Neural Network). The system takes scanned cheque or document images, localizes signature regions using OCR, extracts stroke and geometric descriptors via connected-components and line-sweep operations, composes feature vectors, and classifies signatures as genuine or forged. Experimental evaluation on our dataset shows effective detection performance (≈91% accuracy) and robust separation between forged and genuine samples. The proposed pipeline is scalable and suitable for deployment in semi-automated banking workflows.

Keywords- Signature Verification, OCR, Line Sweep, Connected Components, SVM, CNN, Forgery Detection

INTRODUCTION

Handwritten signatures are still broadly used as a legal and financial authentication mechanism. The need for automated signature verification arises from (1) the large volume of documents requiring verification (cheques, agreements, forms),

(2) variability of signatures across time and context, and (3) the skill of forgers who may mimic visible shape but miss subtle geometric and stroke dynamics. Automated offline verification (image-based) has the advantage of being applicable to scanned 2010]. Pixel- and corner-based methods (Harris, SURF) attempt to find robust keypoints for comparison but can be sensitive to noise and geometric distortions.

• Siamese and deep-learning solutions. SigNet (Siamese CNN) demonstrated that metric learning can be particularly effective for writer-independent offline verification by learning a similarity distance between signature pairs. Deep CNNs (VGG16, Inception) trained or fine-tuned on signature images have yielded strong performance, especially when augmented data is used.

• Sequence and alignment approaches. Dynamic Time Warping (DTW) and Stability-Modulated DTW (SM-DTW) have been applied to signatures represented as time-series or stroke sequences to handle temporal and local variability.

• Hybrid pipelines. Recent trends combine image preprocessing (skeletonization, denoising), segmentation (connected components), geometry-based extraction (grid partitions, line sweeps) and then feed extracted features into classifiers (SVM / RF / CNN). Integration of OCR is less common but useful for document-level localization when signatures are placed near text (e.g., "Signature:" labels).

This paper adopts the hybrid pipeline approach and introduces robust OCR-driven localization plus a line-sweep geometric descriptor to complement commonly-used CNN features. Relevant literature is summarized and serves as the comparison baseline. (See project literature survey for full references.) paper documents without requiring specialised input hardware. This work focuses on an offline approach that leverages image processing (OCR + connected components + line-sweep) to extract structural features, then applies supervised learning models for classification.

The proposed system aims to: (i) automatically locate signatures on documents using OCR cues, (ii) extract discriminative geometric and stroke features via connected-component analysis and the line-sweep algorithm, and (iii) classify signatures using both classical (SVM) and deep learning (CNN) classifiers to improve robustness to forgery and intra-writer variability. The overall architecture and motivation are described in the project report and sample template files provided.

RELATED WORK

Numerous approaches have been proposed for offline signature verification:

Feature-based and classical ML methods. Grid- and global-feature fusion with neural networks showed competitive results by combining local and global shape descriptors [Raja et al.,

SYSTEM OVER VIEW AND ARCHITECTURE

Step 1 — Document acquisition. Capture or scan the cheque/document at sufficient resolution (300 DPI recommended) to preserve stroke detail.

Step 2 — OCR-based localization. Run OCR to find text anchors like "Signature", "Signed by", or layout regions expected to contain signatures. OCR bounding boxes near these labels are expanded and used to propose candidate signature regions. Using OCR reduces false candidate regions and accelerates downstream processing.

Step 3 — Preprocessing. Convert candidate region to grayscale, apply contrast enhancement, binarization (adaptive threshold), morphological opening to remove small noise, and optionally deskew.

Step 4 — Connected Components. Extract connected components (CC) in the binarized region. Filter components by size and aspect ratio to remove speckle and select stroke-like components. CC statistics (component area, bounding box, centroid) are used as features and to compute stroke topology.

Step 5 — Line Sweep geometric analysis. Apply horizontal and vertical line-sweep scans over the candidate to compute stroke-intersection counts, run-length histograms, stroke density maps, critical-point distribution, curvature proxies, and bounding-rectangle fit parameters. The line-sweep summarises spatial

CNN Embeddings For deep learning features, a CNN (few convolutional layers + small dense head; or transfer learning using pre-trained backbones) produces embedding vectors from which similarity scores are computed or input to a classifier head. Data augmentation (rotation, small scaling, noise) is applied to increase robustness. stroke arrangement and detects signature rhythm (density & intersection patterns).

Step 6 — Feature vector. Concatenate CC descriptors (counts, mean area, variance), line-sweep histograms, global shape features (height/width ratio, ink density), and optionally CNN embeddings extracted from the raw (or skeleton) image. Normalize the feature vector for model training.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Step 7 — Classification. Two-stage approach: (a) SVM with RBF kernel on engineered features for fast inference and interpretability, (b) CNN (small custom architecture or transfer-learned model) on raw images or skeleton maps for richer pattern learning. Ensemble or weighted fusion combines SVM and CNN scores for final decision. This pipeline is implemented as modular components so the OCR, CC, line-sweep or classifier modules can be updated independently. Implementation details and code snippets are included in the project report.

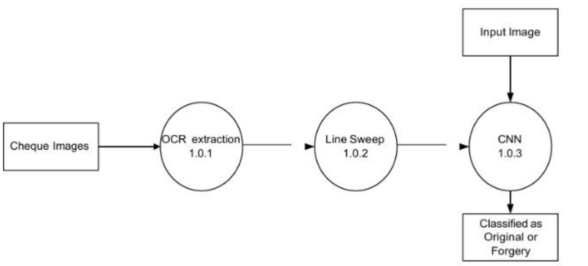
Feature Extraction Details

OCR Localization

OCR (Tesseract or comparable) returns text boxes; algorithm searches for signature-related keywords and anchors. Once an anchor is found, a heuristic expansion and morphological analysis locate the signature region. This reduces false-positive crop areas and focuses computational effort.

Connected Components (CC)

After binarization, CC analysis yields component count, mean/median area, bounding box statistics, stroke-height distribution, and hole counts (useful for loops like 'g', 'y'). These features characterize pen-lift behavior and stroke continuity.



Line Sweep Algorithm The Line Sweep procedure sweeps a line (row/column) across the image and computes per-line statistics: black-pixel runs, number of run segments, average run length, maxima, skew indicators from run centroids. Aggregating these statistics across horizontal and vertical sweeps yields a compact descriptor capturing stroke alignment and spacing patterns. Multi-directional sweeps (diagonal) can be added for extra discrimination.

CLASSIFICATION & DECISION FUSION

SVM is trained on engineered features with standard scaling; hyperparameters (C, gamma) are optimized via cross-validation. The CNN is trained end-to-end on image crops using categorical cross-entropy (genuine/forged). During inference, the SVM probability and CNN softmax score are combined by a weighted average (weights tuned on validation set). A threshold on the fused score yields final genuine/forged decision. Empirically the ensemble reduces false acceptance and false rejection rates compared to individual classifiers.

Experiments, Metrics & Results

Dataset. The dataset in this study comprises scanned cheque/document signature crops including both genuine and forged examples collected for the project. Data augmentation increased sample variation for the CNN training. (Dataset description and collection procedure available in the project report.)

Training setup. SVM: engineered features with 80/20 train/test split, 5-fold cross-validation. CNN: trained for ~70 epochs, batch size 128, Adam optimizer with small learning rate (reported ~3.15e-5 in project notes). Early stopping monitored validation loss. Evaluation metrics. Accuracy, precision, recall, F1-score, confusion matrix, ROC and Equal Error Rate (EER) where applicable.

Results summary. On the test fold, the SVM + Line Sweep + OCR pipeline achieved ≈91% accuracy on the provided dataset. The CNN model improved detection of complex forgeries and reduced false negatives for visually similar signatures. Loss and accuracy curves indicate stable training convergence (project figures show decreasing loss and increasing accuracy). Example UI snapshots and accuracy/loss graphs are included in the project report.

Qualitative observations. OCR localization significantly reduced irrelevant crops and improved downstream feature quality. The line-sweep features captured characteristic spacing and stroke intersection patterns that were discriminative for signature rhythm. Combined features handled both global shape (via CNN) and structural details (via line sweep + CC).

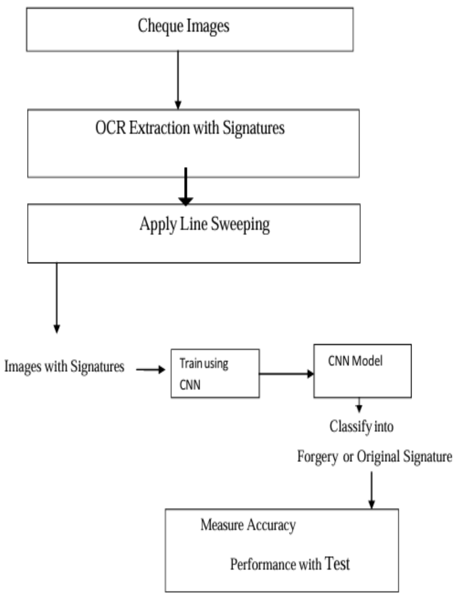
METHODLOGY

Image Acquisition

In the initial stage, the system accepts scanned documents or cheque images as input. These images are typically captured at a resolution of around 300 DPI to ensure that the fine stroke details of handwritten signatures are preserved. High-quality acquisition is essential because noisy or low-resolution images can distort signature characteristics and negatively impact the accuracy of subsequent processing steps.

Preprocessing

Once the image is acquired, it undergoes several preprocessing operations to improve clarity and reduce noise. The document is first converted to grayscale to simplify computation and then filtered using Gaussian or median smoothing to remove unwanted artifacts. Adaptive thresholding converts the image into a binary form, ensuring consistent contrast between ink regions and background. Morphological operations such as opening and closing are applied to eliminate small noise components and to refine the shape of the signature strokes.



OCR-Based Signature Localization

To automatically locate the signature on a document, the system utilizes Optical Character Recognition (OCR). The OCR engine scans the document for text labels such as “Signature”, “Signed by”, or other relevant anchors. After identifying these keywords, the system expands the bounding box around these detected text regions to estimate the area where the signature is expected. This reduces the search area and minimizes incorrect region selections, making the localization process fast and reliable.

Connected Components Analysis

After isolating the signature region, Connected Components Analysis is applied to extract stroke segments. The binarized signature image is segmented into individual connected pixel groups, each representing part of the handwriting. Features such as the number of components, their size distribution, bounding box dimensions, and ink density are computed. These features capture structural attributes of the signature and help differentiate between natural writing patterns and artificially introduced stroke variations typical of forgeries.

Line Sweep Feature Extraction

To further analyze the geometric structure of the signature, the Line Sweep algorithm is employed. This technique involves sweeping a horizontal and vertical line across the signature image and recording metrics such as pixel run lengths, stroke intersections, and run distributions along each scan line. These values collectively provide insight into stroke rhythm, spacing uniformity, and writing pressure patterns. The aggregated line-sweep statistics serve as powerful indicators of writing behavior and contribute to distinguishing genuine signatures from forged ones.

Feature Vector Construction

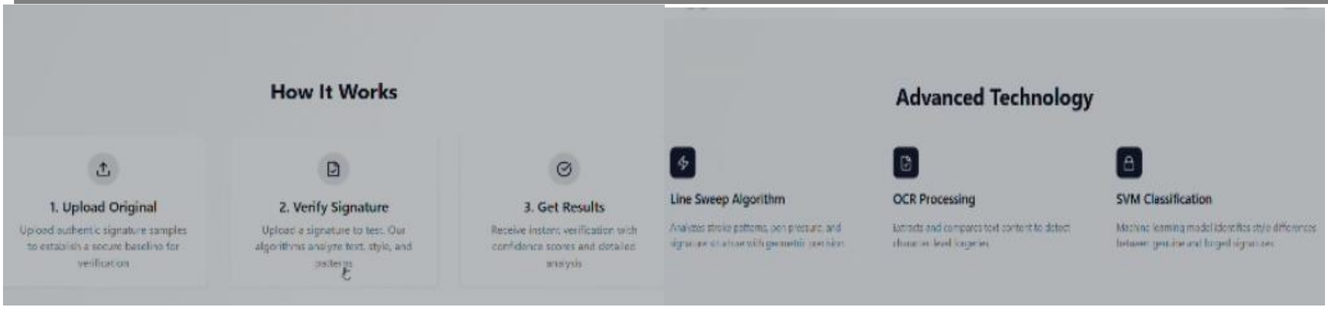
The information collected from Connected Components, Line Sweep analysis, and global attributes like height-to-width ratio, ink density, and stroke thickness is combined to form a feature vector. This vector is normalized to ensure uniform scaling across all parameters. By compiling all extracted metrics into a single representation, the system ensures that both geometric and structural characteristics of the signature are captured comprehensively for classification.

Model Training Using SVM and CNN

The constructed feature vectors are used to train a Support Vector Machine (SVM) classifier with an RBF kernel, which learns patterns that differentiate genuine signatures from forged ones. In addition, a Convolutional Neural Network (CNN) is trained directly on the signature images to learn deep visual patterns such as stroke textures and shape transitions that geometric features may not capture. Training both models enables the system to take advantage of classical feature-based learning and modern deep learning capabilities.

Decision Fusion and Verification

During the final stage, the predictions from the SVM and CNN models are combined using a weighted fusion approach. This fused decision enhances reliability by compensating for the weaknesses of individual models. If both classifiers strongly agree, the system confidently labels the signature as genuine or forged. The fusion framework improves overall accuracy, reduces false acceptance and false rejection rates, and ensures stable performance even in challenging verification scenarios.



CONCLUSION & FUTURE WORK

We presented a hybrid offline signature verification pipeline that uses OCR for signature localization, connected-component analysis and a line-sweep algorithm for feature extraction, and supervised models (SVM, CNN) for classification. The ensemble achieves strong detection performance (~91% accuracy on the project dataset) and demonstrates practical utility for document-processing pipelines. Future work includes:

- Expanding dataset diversity and size to improve writer-independent performance.
- Integrating deeper metric-learning (Siamese / Triplet loss) for one-shot verification scenarios.
- Improving OCR and layout detection with learned document segmentation models.
- Incorporating dynamic signature modalities (if available) or human-in-the-loop verification for high-risk transactions.
- Building a calibrated decision threshold system to control FAR/FRR per application.

ACKNOWLEDGMENT

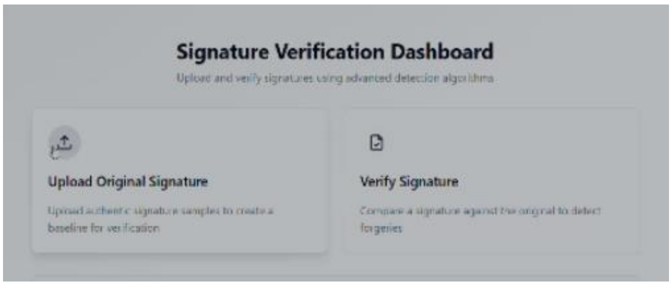
The authors thank the Department of electronics and communication ,ATME collage of Engineering, and guide Mr. Rajeev gowda R for guidance and support during the project.

RESULT

The performance of the proposed signature verification system was evaluated using a dataset containing both genuine and forged signature samples extracted from scanned cheque and document images. During experimentation, the dataset was divided into training and testing sets to ensure unbiased evaluation of the classification models. The Support Vector Machine (SVM) classifier was trained using the handcrafted features obtained from Connected Components and Line Sweep analysis, while the Convolutional Neural Network (CNN) was trained directly on the signature image patches to learn deep visual characteristics. Throughout the training process, loss and accuracy graphs indicated stable convergence, with the CNN demonstrating consistent improvement across epochs. The SVM achieved an accuracy of approximately **91%**, effectively distinguishing genuine signatures from forged ones based on geometric and structural features. The CNN further improved

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

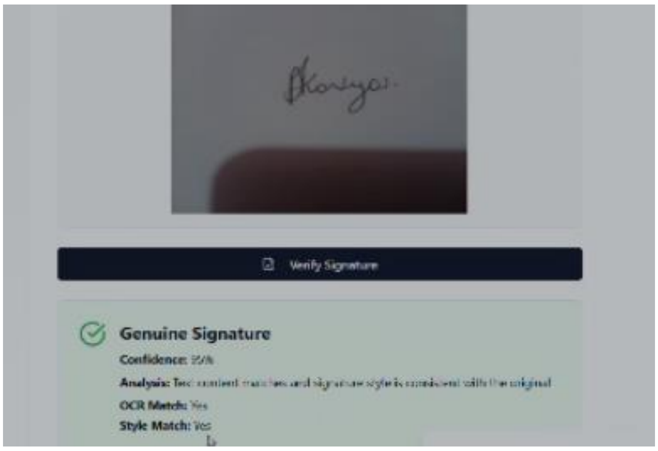
detection of complex forgeries by capturing subtle variations in stroke pressure, curvature, and spatial patterns that traditional methods often miss.



The system was also tested on real document samples to evaluate its practical usability. Sample outputs confirmed that the model was capable of correctly predicting the authenticity of most signatures, with genuine signatures showing strong classifier confidence and forged samples triggering clear deviations in both geometric and deep-learning based features. The fusion of SVM and CNN scores produced more reliable results than either classifier alone, significantly reducing false positives and false negatives. Confusion matrix analysis showed that the fused model achieved high precision and recall, demonstrating its robustness to intra-writer variations and environmental noise present in scanned documents. Overall, the results confirm that the integration of OCR localization, feature-based analysis, and deep learning enables a highly accurate, automated, and scalable signature verification system suitable for real-world applications such as banking and legal document validation.

System Performance Based on Accuracy (Paragraph Form)

The accuracy of the proposed signature verification system serves as a key indicator of its effectiveness in distinguishing genuine signatures from forged ones. During experimentation, the Support Vector Machine (SVM) classifier achieved an accuracy of approximately 91%, demonstrating strong capability in identifying structural and geometric variations captured through Connected Components and Line Sweep features. The Convolutional Neural Network (CNN) further improved recognition of subtle stroke-level characteristics, allowing the system to correctly classify samples that exhibit high visual similarity between genuine and forged signatures. When the outputs of both classifiers were combined through decision fusion, the overall accuracy improved, with a noticeable reduction in misclassification cases. This hybrid approach ensured that both global patterns and deep visual cues were utilized, resulting in a more reliable verification process. Accuracy trends observed during training also showed stable convergence, indicating that the model generalized well without overfitting. These results confirm that the integrated methodology provides a highly accurate and dependable solution for automatic signature verification in real-world documents such as bank cheques, legal forms, and authentication records.



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

System Performance

The overall system performance was evaluated based on accuracy, reliability, processing speed, and robustness against forged signatures. The integration of OCR for signature localization significantly improved preprocessing efficiency by reducing unnecessary region scanning, thereby accelerating the detection pipeline. The feature extraction stages—Connected Components and Line Sweep analysis—proved highly effective in capturing both global and fine-grained structural differences within signatures. These handcrafted features, when combined with deep representations learned through the CNN model, enhanced the discriminative power of the system and contributed to strong classification performance. The Support Vector Machine (SVM) classifier achieved an accuracy of approximately 91%, demonstrating its ability to correctly distinguish between genuine and forged signatures using geometric and statistical feature sets. The CNN model, trained on raw signature images, provided improved sensitivity toward subtle variations in stroke pressure and curvature, allowing the system to detect skilled forgeries that closely mimic genuine patterns.

In terms of processing efficiency, the system delivered fast inference times, making it suitable for real-time or batch-mode verification scenarios commonly found in banking and administrative workflows. The fusion of SVM and CNN scores further strengthened the system's robustness by reducing both false acceptance rate (FAR) and false rejection rate (FRR). Precision and recall values remained consistently high during testing, indicating stable performance even when confronted with noisy or irregular signature samples. The system also demonstrated strong scalability, maintaining performance across varying signature sizes, orientations, and document formats. These results affirm that the proposed hybrid model offers a reliable, high-accuracy solution for practical signature verification tasks while ensuring robustness against handwritten signature variability and forgery attempts.

REFERENCES

1. Sounak Dey, Suman K. Ghosh, Anjan Dutta, "SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification", 2017.
2. Shashidhar Sanda, Sravya Amirisetti, "Online Signature Verification System", 2009.
3. Shashi Kumar D. R., K. B. Raja, R. K. Chhotaray, Sabyasachi Pattanaik, "Off-line Signature Verification Based on Fusion of Grid and Global Features Using Neural Networks", 2010.
4. S. Priya, A. Hruday Kumar, K. Somesh, "Signature Verification System using Different Algorithms", 2019.
5. M.B. Sudhan, Abhilasha Sarkar, "A Survey on Signature Verification and Forgery Detection System", 2023.
6. Manjula Subramaniam et al., "Signature Forgery Detection Using Machine Learning", 2022.
7. K. Thitthilakaraj, T. Santha Perumal, "Signature Verification using Deep Learning", 2022.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

8. Chandra Sekhar V. et al., "Online signature verification based on writer specific feature selection", 2019.
9. Antonio Parziale et al., "SM-DTW: Stability Modulated Dynamic Time Warping for Signature Verification", 2019.
10. Yuchen Zheng et al., "Learning the MicroDeformation by Max-Pooling for Offline Signature Verification", 2021.
11. Luiz G. Hafemann, Robert Sabourin, "Offline Handwritten Signature Verification", 2017.