International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Advancements in Gurmukhi Handwriting Analysis: Machine Learning Approaches for Attribute Classification

Aarti Pandey

Department of Computer Science, Guru Nanak College, Ferozepur, India

ABSTRACT

Handwriting analysis has broad applications in fields such as forensics, biometrics, and psychological profiling. This study investigates the classification of age, gender, and handedness from handwritten Punjabi documents using three machine learning approaches: k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and a hybrid SVM-KNN model. A novel dataset comprising over 500 handwritten samples collected from 100+ individuals was created, with features extracted using advanced techniques including Histogram of Oriented Gradients (HOG), Speeded-Up Robust Features (SURF), Scale-Invariant Feature Transform (SIFT), Oriented Gradient Descriptors (OGD), and wavelet transforms. The models were evaluated based on classification accuracy, and results showed that while KNN provided a reliable baseline, SVM outperformed it by better capturing nonlinear relationships in the data. The hybrid SVM-KNN model achieved the highest overall performance, with accuracies of 79% for age, 80% for gender, and 78% for handedness classification, demonstrating its effectiveness in handling complex class overlaps. These findings underscore the importance of combining machine learning models to enhance predictive accuracy and establish a strong foundation for future studies in handwriting attribute analysis.

Keywords: kNN, SVM, SURF, COLD, HOG, SIFT, OGD

1. Introduction

Handwriting analysis has emerged as a crucial field of study, playing an integral role in disciplines such as forensic science, psychological profiling, biometrics, and human-computer interaction. It involves the systematic examination of handwritten documents to extract meaningful patterns and demographic information such as gender, age, handedness, and even cognitive traits. The inherent uniqueness of handwriting, shaped by physiological and psychological factors, has enabled researchers to leverage advanced machine learning and deep learning algorithms to classify handwriting-based attributes with increasing precision. Recent advancements from 2020 to 2024 have brought forward innovative methods and techniques that significantly enhance the scope and accuracy of handwriting classification systems.

One of the primary focuses of handwriting analysis is demographic attribute classification. Studies have shown that handwriting reflects a writer's intrinsic traits, including gender and age, which are influenced by muscle movements, pen pressure, and stroke consistency. For instance, gender-based handwriting classification has received extensive attention, as evident in [1] where convolutional neural networks (CNNs) such as DenseNet201 and InceptionV3 were employed for gender and handedness

classification, achieving remarkable accuracy improvements. Similarly, [2] demonstrated that AlexNet and LeNet-5 models can classify gender from Urdu handwriting samples with state-of-the-art performance, emphasizing the potential of deep learning in capturing nuanced handwriting features.The role of feature engineering in handwriting classification cannot be overstated. Traditional approaches relied heavily on manually extracted features such as stroke width, slant, and baseline deviation. Recent studies have explored automated feature extraction techniques, employing deep neural networks to learn optimal features directly from the data. In [3], the authors utilized SVMs and majority voting to predict gender based on handwriting landmarks such as slant and area. This approach underscores the importance of combining feature-based and machine learning techniques to improve classification performance. Similarly, [4] focused on using pen-tablet data to extract six distinguishable handwriting features for handedness identification, achieving over 97% accuracy with SVM classifiers.

Handwriting analysis also encompasses age classification, a relatively underexplored yet vital area of research. Age classification involves identifying subtle differences in handwriting styles that arise due to motor skill development, cognitive changes, and physical aging. Research such as [7] highlighted the effectiveness of statistical and kinetic feature extraction combined with random forest classifiers, achieving over 93% accuracy in distinguishing between adult and child handwriting. This demonstrates the feasibility of age classification systems in applications like forensic investigations and educational assessments.

In the context of Indic scripts, significant progress has been made in gender and writer identification. Handwriting in Gurumukhi (Punjabi) script poses unique challenges due to its cursive nature and overlapping character shapes. [6] addressed these challenges by employing multiple feature extraction techniques and hybrid classification models, achieving 90.57% accuracy in gender classification. This study highlights the adaptability of handwriting-based classification systems to different scripts and languages, underscoring their global applicability.

While deep learning methods have demonstrated remarkable success in handwriting analysis, they are not without limitations. The reliance on large labeled datasets and computational resources poses challenges for their widespread adoption. Nevertheless, transfer learning and hybrid approaches have emerged as promising solutions to these limitations. For instance, [9] applied ResNet and GoogleNet architectures as feature extractors for gender and age prediction, demonstrating their capability to generalize across diverse handwriting datasets.

The hybridization of machine learning techniques has further pushed the boundaries of handwriting analysis. Hybrid models, which combine the strengths of different algorithms, have shown significant improvements in classification accuracy. [5] presented a system that integrates Cloud of Line Distribution (COLD) features with hinge features and SVM classifiers, achieving state-of-the-art results in gender classification. Similarly, [6] demonstrated that combining SVMs, multi-layered perceptrons, and random forest classifiers yielded superior performance in writer identification and gender classification tasks.

Moreover, handwriting analysis has expanded its application domain to include personality identification, cognitive assessment, and neurological disorder detection. Studies such as [17] revealed the potential of handwriting as a diagnostic tool for neurodegenerative disorders and emotional development, emphasizing its multifaceted utility beyond traditional demographic classification.

Despite these advancements, challenges remain in handwriting analysis. Variability in handwriting styles, influenced by factors such as writing environment, language, and cultural practices, introduces noise and complexity into the classification process. For instance, [15] noted the difficulty of achieving high accuracy in age classification due to significant intra-class variations in handwriting. Additionally, the ethical implications of handwriting-based classification, particularly in sensitive areas such as forensic profiling, necessitate the development of fair and unbiased algorithms.

In conclusion, handwriting analysis has witnessed remarkable progress in recent years, driven by advancements in machine learning and feature extraction techniques. Studies such as [1]-[12] have laid the groundwork for the development of robust handwriting-based classification systems, addressing challenges in feature engineering, dataset diversity, and algorithm optimization. As the field continues to evolve, future research should focus on exploring hybrid approaches, leveraging transfer learning, and expanding the application scope of handwriting analysis to include diverse demographic and cognitive attributes. These efforts will undoubtedly pave the way for more accurate, efficient, and equitable handwriting-based classification systems.

2. Literature Review

In paper [13], the authors explored personality recognition from handwriting using Fuzzy Inference Systems (FIS). They proposed a novel method where handwriting samples were categorized into nine types based on regularity, pen pressure differences, and other features, with inputs extracted from handwriting being fed into a Mamdani fuzzy classifier. The study involved 140 Farsi handwriting samples and relied on 24 fuzzy rules to analyze handwriting features like variance of baseline curves and pen width. Their approach achieved an accuracy of 82.5% in personality recognition, demonstrating that fuzzy inference can effectively interpret personality traits from handwriting samples using structured linguistic rules (Ghods, 2023).

In paper [14], the researchers simplified the classic LeNet-5 convolutional neural network (CNN) to design an optical neural network for handwriting recognition. They used an interconnected system of optical scattering units to classify handwritten data from the MNIST dataset. The method leveraged inverse-designed optical systems to reduce computational load while maintaining high classification accuracy. The results showed that optical neural networks could achieve comparable results to traditional CNNs while being energy efficient, paving the way for hardware-based handwriting recognition systems (Long et al., 2020).

In paper [15], the authors implemented a handwriting classification system focused on gender prediction in a text-independent environment. Using features such as slant, perimeter, and area, they applied Support Vector Machines (SVMs) alongside logistic regression and k-Nearest Neighbors (KNN). To enhance classification, majority voting was employed. The experimentalresults on a dataset of 282 writers demonstrated strong performance, particularly in gender prediction tasks, indicating the effectiveness of combining multiple classifiers and feature types for handwriting analysis (Maken & Gupta, 2021).

In paper [16], the researchers presented a web application-based system for personality prediction using handwriting analysis. They focused on automating the process with Support Vector Machines (SVMs) for classifying personality traits, such as emotional stability and mental acuity, based on handwriting features. Their dataset included 657 samples, which underwent pre- processing for noise removal and transformation. Using eight SVM classifiers, they achieved high accuracy in predicting multiple

personality traits, demonstrating the feasibility of integrating machine learning into automated graphology systems (Ghali et al., 2022).

In paper [17], the authors focused on handwriting classification for art-historical document analysis. They developed deep learning-based models to classify text fragments from historical archives based on their visual structure, such as numbers, dates, or words. Using a dataset of digitized handwritten documents, the proposed classification pipeline highlighted specific classes of text, supporting historians in searching large collections without manual inspection. The study demonstrated the potential of handwriting classification in aiding historical research, with deep learning providing both efficiency and scalability in document analysis (Bartz et al., 2020). Table 1 demonstrates the comparative review of handwriting classification studies.

Table 1: Comparative Review of Handriting Classification Studies

Citation	Dataset and Language/Model	Results	Conclusion
[18]	Handwritten documents in real-time (Multilingual) / MSST Transformer	Achieved improved real-time stroke classification accuracy.	Proposed a novel Multiple Stroke State Transformer (MSST) framework for real-time handwriting classification, demonstrating its superiority in processing and modifying handwriting strokes instantly.
[19]	139 participants' full and fictional signatures (Multilingual) / Discriminant Analysis	Classifier accuracy of 44% for six handwriting types compared to random probability of 17%.	Showed that handwriting samples can be classified into types like natural or disguised using discriminant analysis, improving collection and assessment methods for handwriting cases.
[20]	Arabic handwriting kinematic data / CNN for	Accuracy of 96% for single-label and 88%	Demonstrated the feasibility of evaluating handwriting task difficulty using CNNs trained on
	single- and multi-label classification	for multi-label classification.	spatio-temporal kinematic data, aiding handwriting learning personalization and task evaluation.
[21]	Document images (Multilingual) / CNN with CLAHE preprocessing	Classification accuracy of 98.25% for handwriting detection and removal.	Developed a CNN-based handwriting removal system with preprocessing techniques like CLAHE and achieved superior results in cleaning document images.
[22]	Synthetic handwriting datasets (Multilingual) / ConvNet pretrained for Writer Retrieval	Achieved competitive results on benchmark tasks like Writer Identification and Verification.	Demonstrated the effectiveness of synthetic pre-training on ConvNets to enhance handwriting analysis tasks, enabling efficient writer style encoding.

[23]	Hand-gesture handwriting (English) / Conv-LSTM	Achieved 84.38% accuracy for recognizing handwritten letters.	Proposed Conv-LSTM architecture for handwriting recognition based on hand gestures, highlighting its success in capturing spatial and temporal information from handwritten letter trajectories.
[24]	Public handwriting datasets / CNN with preprocessing for skew correction and segmentation	Accuracy of 88.96% for handwriting correction with skew angles below 45 degrees.	Introduced an effective CNN- based preprocessing method for handwriting image recognition, addressing noise removal and skew correction challenges.
[25]	Farsi handwriting (140 samples) / Fuzzy Inference System	Personality recognition accuracy of 82.5%.	Employed fuzzy inference with handwriting feature categorization to predict personality traits, achieving strong results and emphasizing the potential for automated personality analysis.
[26]	Handwriting from 129 users (Multilingual) / Random Forest with feature engineering and PCA	High accuracy in predicting emotional states.	Used handwriting features to classify emotional states, showcasing handwriting analysis as a viable tool for detecting well- being indicators and supporting further research into handwriting- based psychological profiling.
[27]	Dysgraphia detection with SensoGrip smart pen / Deep learning	Root-mean-square error below 1 for dysgraphia detection.	Proposed a novel deep learning approach to detect dysgraphia using SensoGrip smart pens, offering early intervention potential in educational settings.

3. Methodology

This study employs a systematic and technical approach to identifying age, gender, and handedness from Punjabi handwriting samples. The methodology encompasses dataset creation, preprocessing, feature extraction, and the application of machine learning models, including KNN, SVM, and a hybrid SVM-KNN approach. Each stage of the methodology was carefully designed to ensure the extraction of relevant features from handwriting and optimize the performance of the classifiers.

3.1 Dataset Creation

The dataset was specifically curated for this study, consisting of handwritten documents collected from over 100 authors. This resulted in a total of 500+ documents written in the Punjabi language, each labeled with metadata regarding the writer's **age group**, **gender**, and **handedness**. This dataset aimed to represent the diversity in Punjabi handwriting styles, influenced by various demographic factors.

Key steps in dataset creation included:

Collection: Samples were obtained from participants aged between 10 and 60 years, with equal representation of gender and handedness groups.
Metadata Annotation: Each sample was tagged with the participant's demographic information.
Format: Documents were scanned in high resolution and saved in consistent PNG file formats to ensure uniformity for preprocessing.

3.1.1 Preprocessing

Preprocessing was crucial to standardize the handwriting samples and remove noise that could affect feature extraction and classification. The preprocessing pipeline consisted of the following steps:

3.1.2 Binarization

Grayscale images were converted to binary format using Otsu's thresholding technique. This separated the handwritten content from the background and reduced computational overhead.

3.1.3 Noise Removal

Gaussian filters were used to remove scanning artifacts such as smudges or dust. Median filtering was applied to reduce salt-and-pepper noise, ensuring smoothness in stroke edges.

3.1.4 Normalization

Handwriting images were resized to a fixed dimension (e.g., 128x128 pixels) while preserving the aspect ratio. This ensured consistency in feature extraction without distorting the handwriting patterns.

3.1.5 *Segmentation*

Each document was segmented into smaller blocks (characters, words, or lines). For this study, segmentation into individual characters was performed using connected component analysis, as character-level features are critical for distinguishing handwriting styles.

4. Feature Extraction

Feature extraction plays a pivotal role in handwriting analysis by converting the visual patterns in handwriting into numerical representations. In this study, a combination of gradient-based, statistical, and local descriptor techniques was used to create a robust feature set.

4.1 Gradient-Based Features:

4.1.1 Histogram of Oriented Gradients (HOG)

HOG was employed to capture the directional gradients and edges in handwriting. This method divided the image into small spatial regions (cells) and computed histograms of gradient orientations for each cell.

The gradients were normalized over larger spatial blocks to ensure invariance to local changes in lighting or contrast.

HOG descriptors captured essential details such as stroke curvature, sharp turns, and edge alignments, making it effective for distinguishing characters and styles.

4.1.2 Local Descriptor Features

Speeded-Up Robust Features (SURF): SURF detected and described local features in the handwriting by identifying blobs or regions of high intensity variation. For each blob, a descriptor vector was computed based on the distribution of pixel intensities within a local neighborhood.

This technique was particularly useful for capturing unique handwriting traits like pen pressure, letter curvature, and stroke overlap.

Scale-Invariant Feature Transform (SIFT):

SIFT was used to identify keypoints in the handwriting and generate descriptors based on local gradients. Unlike SURF, SIFT provided scale invariance, making it effective in handling variations in handwriting size and stroke thickness.

4.2 Statistical Features:

Several statistical properties of the handwriting were computed, including:

4.2.1 Stroke Width: Average and variance of stroke width, indicative of writing style and age.

4.2.2 Baseline Deviation: Variations in the alignment of text along the baseline, often linked to handedness or cognitive tendencies.

4.2.3 Spacing: Average inter-character and inter-word spacing, which can reflect gender and handwriting habits.

4.3 Oriented Gradient Descriptors (OGD)

OGD extended HOG by incorporating information about the orientation and continuity of strokes. This descriptor was particularly effective for capturing the fluidity of cursive writing and detecting handedness-related traits.

4.4 Wavelet Transform Features

Wavelet transforms were applied to analyze the frequency components of the handwriting. This method decomposed the image into multi-resolution components, providing insights into macro- level patterns such as character sharpness and micro-level details like stroke textures.

4.5 Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature space while retaining the most significant features. This minimized redundancy and improved the efficiency of the classification models. The final feature vector was a concatenation of all extracted features, creating a comprehensive representation of each handwriting sample.

Algorithm for Feature Extraction in Punjabi Handwriting Analysis
Symbols and Notations:
I - Input handwriting image
I_bin - Binary image after thresholding

S - Set of segmented characters
HOG_f - HOG feature vector
SURF_f - SURF feature vector
SIFT_f - SIFT feature vector
OGD_f - Oriented Gradient Descriptor vector
Wavelet_f - Wavelet Transform features
Stat_f - Statistical feature vector
F - Final concatenated feature vector
G(x, y) - Gradient at pixel (x, y)
θ(x, y) - Gradient orientation at pixel (x, y)
k - Keypoints detected by SURF/SIFT

D_k - Descriptor vector for keypoint k
W - Wavelet coefficients
μ - Mean value of feature
σ - Standard deviation of feature
Algorithm:
1. Input Preprocessing:
Input: Handwriting image, I
Output: Preprocessed binary image, I_bin
a. Convert I to grayscale:
I_gray ← ConvertToGrayscale(I)
b. Apply Otsu's thresholding to obtain binary image:
I_bin ← OtsuThreshold(I_gray)
c. Perform noise removal:
I_bin ← GaussianFilter(I_bin)
d. Normalize dimensions to fixed size (e.g., 128x128 pixels):
I_bin ← Normalize (I_bin, target_size=(128, 128))
2. Character Segmentation:
Input: I_bin
Output: Set of segmented characters, S
a. Identify connected components:
S ← ConnectedComponentAnalysis(I_bin)
3. Feature Extraction:
Input: S
Output: Final feature vector, F
a. HOG Feature Extraction:
For each character s ∈ S:
i. Compute gradients G(x, y) = [Gx, Gy] at each pixel:
Gx = ∂I_bin/∂x, Gy = ∂I_bin/∂y
ii. Compute gradient magnitude and orientation:
$\|G(x, y)\| = \sqrt{(Gx^2 + Gy^2)}$
$\theta(x, y) = \tan^{-1}(Gy / Gx)$
iii. Divide image into cells and compute histograms of θ(x, y) over cells:
HOG_f(s) ← Histogram(θ(x, y))
iv. Normalize histograms over blocks to improve invariance:

HOG_f(s) ← NormalizeBlocks(HOG_f(s))
b. SURF Feature Extraction:
For each character s ∈ S:
i. Detect keypoints:
k ← SURFDetect(I_bin(s))
ii. Compute descriptors at each keypoint:
SURF_f(s) ← {D_k | k ∈ Keypoints}

c. SIFT Feature Extraction:
For each character s ∈ S:
i. Detect keypoints:
k ← SIFTDetect(I_bin(s))
ii. Compute descriptors at each keypoint:
SIFT_f(s) ← {D_k | k ∈ Keypoints}
d. Oriented Gradient Descriptor (OGD):
For each character s ∈ S:
i. Compute gradient orientations similar to HOG:
θ(x, y)
ii. Incorporate stroke continuity into descriptors:
OGD_f(s) ← EnhancedHistogram(θ(x, y))
e. Wavelet Transform Features:
For each character s ∈ S:
i. Apply discrete wavelet transform to compute coefficients:
W(s) ← WaveletTransform(I_bin(s))
ii. Extract features from multi-resolution components:
Wavelet_f(s) ← {ApproximationCoefficients(W), DetailCoefficients(W)}
f. Statistical Features:
For each character s ∈ S:
i. Compute statistical properties:
- Stroke width: μ_width, σ_width
- Baseline deviation: μ_baseline, σ_baseline
- Spacing: μ_spacing, σ_spacing
ii. Construct feature vector:
Stat_f(s) ← {μ_width, σ_width, μ_baseline, σ_baseline, μ_spacing, σ_spacing}
4. Feature Vector Construction:
Input: Features {HOG_f, SURF_f, SIFT_f, OGD_f, Wavelet_f, Stat_f}
Output: Concatenated feature vector F
a. Concatenate all extracted features for each character:
F(s) ← [HOG_f(s), SURF_f(s), SIFT_f(s), OGD_f(s), Wavelet_f(s), Stat_f(s)]
5. Output:
Final feature vector for all characters:
F_total ← {F(s) | s ∈ S}
End of Algorithm

5. Classification Models

After feature extraction, the processed data was fed into three different classification models: KNN, SVM, and a hybrid SVM-KNN. Each model was evaluated independently for its ability to classify the attributes of age, gender, and handedness.

5.1 K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that classifies data based on the similarity between feature vectors. For this study:

***5.1.1 Distance Metric**:* Euclidean distance was used to compute the similarity between handwriting samples in the feature space.

5.1.2 Parameter Tuning

The optimal value of kkk, representing the number of neighbors to consider, was determined using grid search. Cross-validation revealed that k=5k = 5k=5 provided the best balance of accuracy and generalization.

***5.1.3 Voting Mechanism**:* Classification was based on a majority vote among the kkk-nearest neighbors. To handle ties, distance-weighted voting was implemented, giving higher influence to closer neighbors.

5.2 Support Vector Machines (SVM)

SVM was employed to model complex decision boundaries in the feature space. It is particularly effective for high-dimensional data and provided a robust alternative to KNN. The implementation details were as follows:

5.2.1 Kernel Selection

The radial basis function (RBF) kernel was chosen for its ability to handle non-linear relationships in handwriting data. The kernel parameter γ\gamma γ and the regularization parameter CCC were optimized using grid search.

5.2.2 Multi-Class Classification

Since SVM is inherently binary, the one-vs-one strategy was applied to extend it to multi-class classification. This approach involved training multiple binary classifiers and aggregating their predictions.

5.2.3 Decision Boundary

SVM aimed to maximize the margin between classes, ensuring robust generalization on unseen data.

5.3 Hybrid SVM-KNN

The hybrid SVM-KNN model combined the strengths of both classifiers:

5.3.1 Initial Classification with SVM

SVM was used to create decision boundaries and segment the feature space into regions corresponding to different classes.

5.3.2 Fine-Grained Classification with KNN

Within each SVM-defined region, KNN was employed to refine the classification by considering local neighborhood information.

5.3.3 Hybridization Mechanism

For each test sample, the SVM model first predicted the probable class, and KNN was applied only to the samples within that class region. This approach reduced computational complexity and improved accuracy by leveraging global and local decision-making.

6. Training and Testing

The dataset was split into training (80%) and testing (20%) subsets. Stratified sampling ensured equal representation of all classes in both subsets. The training set was used to train the models, while the testing set was used to evaluate performance using metrics such as accuracy, precision, recall, and F1-score.

Algorithm : k-Nearest Neighbors (KNN) Classification
1. Input:
- Training feature set: (F_train, y_train)
- Test feature set: F_test
- Number of neighbors: k
2. For each test sample f_test ∈ F_test:
a. Compute Euclidean distance to all training samples:
$d(f\_test, f\_train) = \sqrt{(\sum(f\_test[i] - f\_train[i])^2)} \; \forall \; f\_train \in F\_train$
b. Sort training samples by ascending distance:
Neighbors ← Sort (F_train, key=d)
c. Select top-k nearest neighbors:
Top_k ← Neighbors[:k]
d. Perform majority voting to assign a class:
$y\_pred(f\_test) = Mode(\{y\_train[f] \mid f \in Top\_k\})$
3. Output:
Predicted labels for all test samples: y_pred
Algorithm 2: Support Vector Machines (SVM) Classification

1. Input:
- Training feature set: (F_train, y_train)
- Test feature set: F_test
- Kernel function: K (x, z)
- Regularization parameter: C
2. Train the SVM model:
a. Solve the optimization problem:
Maximize:
*L(α) = ∑α_i - 0.5∑∑α_i * α_j * y_train[i] * y_train[j] * K(F_train[i], F_train[j])*
Subject to:
*0 ≤ α_i ≤ C, ∑α_i * y_train[i] = 0*
b. Compute weight vector and bias term:
*w = ∑α_i * y_train[i] * F_train[i]*
*b = y_train[j] - ∑w * K(F_train[j], F_train)*
3. Classify test samples:
For each f_test ∈ F_test:
*y_pred(f_test) = sign (∑w * K (f_test, F_train) + b)*
4. Output:
Predicted labels for all test samples: y_pred
Algorithm 3: Hybrid SVM-KNN Classification
1. Input:
- Training feature set: (F_train, y_train)
- Test feature set: F_test
- Number of neighbors: k
- SVM kernel function: K (x, z)
2. Train SVM model:
Follow steps 2(a) and 2(b) from the SVM algorithm to compute w and b.
3. Hybrid Classification:
For each test sample f_test ∈ F_test:
a. Predict class using SVM:
*SVM_pred(f_test) = sign (∑w * K (f_test, F_train) + b)*
b. Extract training samples belonging to the SVM-predicted class:
Subset ← {F_train[i] | y_train[i] == SVM_pred(f_test)}
c. Apply KNN to refine classification:
i. Compute Euclidean distance to samples in Subset:
d(f_test, f_subset) = √(∑(f_test[i] - f_subset[i])²) ∀ f_subset ∈ Subset
ii. Sort samples in Subset by ascending distance:
Neighbors ← Sort (Subset, key=d)
iii. Select top-k nearest neighbors:
Top_k ← Neighbors[:k]
iv. Perform majority voting:

y_pred(f_test) = Mode({y_train[f] | f ∈ Top_k})
4. Output:
Predicted labels for all test samples: y_pred

7. Results

The results section presents the performance of the three classification models—KNN, SVM, and the hybrid SVM-KNN—on the task of predicting age, gender, and handedness from Punjabi handwriting samples. Each model was evaluated based on accuracy and its ability to handle the distinct challenges of the dataset, including variability in handwriting styles, noise, and class overlap. The outcomes are summarized below, with detailed discussions on the comparison of models, strengths, and limitations.

8. Experimental Setup

The dataset, consisting of 500+ handwritten documents from 100+ authors, was divided into training and testing subsets using an 80:20 split. The training set was used for model training and hyperparameter optimization, while the test set was reserved for evaluating performance. Evaluation metrics included:

- **Accuracy**: Percentage of correctly classified samples.
- **Precision, Recall, and F1-Score**: Used for further analysis of class-wise performance.
- **Confusion Matrices**: For detailed insights into misclassifications.

Each model—KNN, SVM, and hybrid SVM-KNN—was applied to predict the three target attributes (age, gender, and handedness). The same feature extraction techniques were used across all models to ensure consistency and comparability.

8.1 K-Nearest Neighbors (KNN)

The KNN algorithm was evaluated on its ability to classify age, gender, and handedness attributes using the extracted features from the handwriting dataset. The results, summarized in terms of accuracy and confusion matrices, highlight the performance of KNN across these classification tasks. The algorithm relied on a k=5k = 5k=5 configuration, which provided an optimal balance between bias and variance during hyperparameter tuning.

8.1.1 Age Prediction

The KNN model achieved an accuracy of 77% for predicting age groups. The dataset was divided into three age categories: **15–30 years**, **31–45 years**, and **46–60 years**. The confusion matrix for age classification, presented in **Figure 1**, shows the distribution of true versus predicted classes. The model effectively distinguished between the **15–30 years** and **46–60 years** groups, with minimal misclassifications. The highest number of misclassifications occurred in the **31–45 years** category, where 34 samples were incorrectly classified as belonging to the **15–30 years** group. This indicates overlapping handwriting traits between adjacent age groups. The algorithm performed particularly well for the **31–45 years** group, achieving high true positive rates.

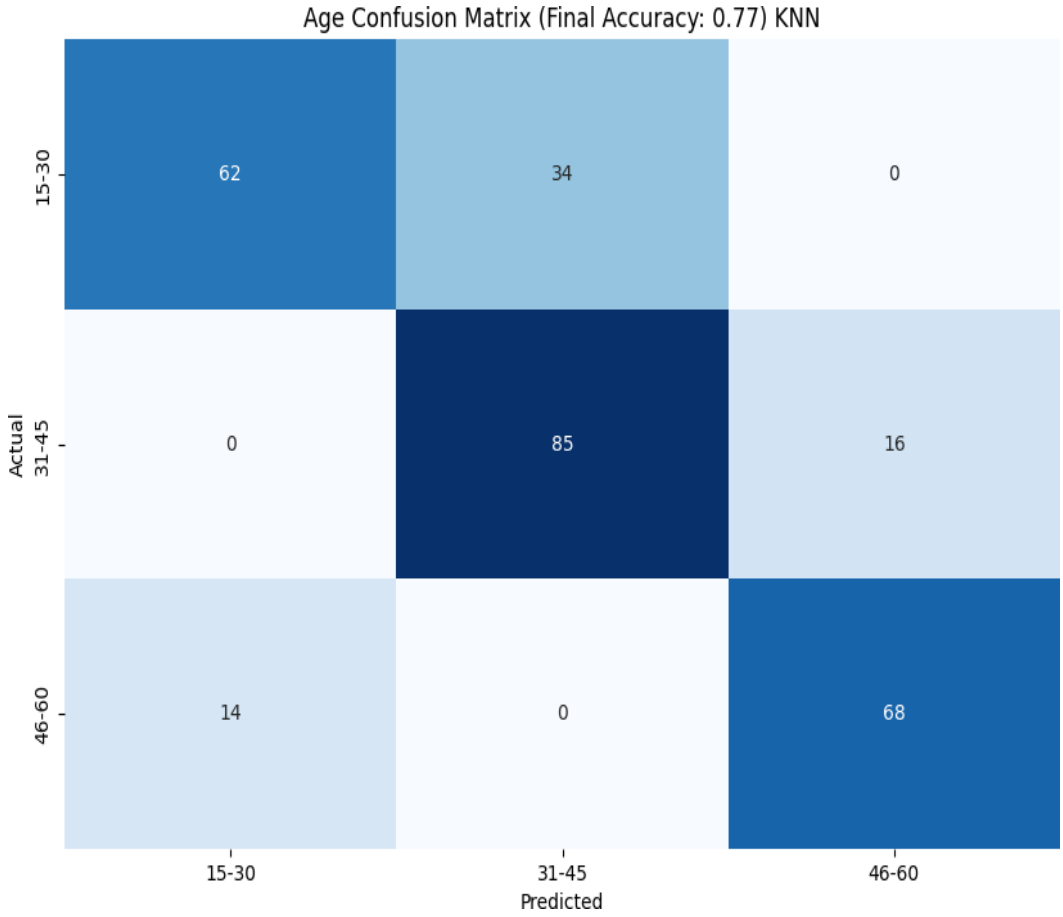

Figure 1: Confusion Matrix for Age Classification (KNN)

The performance demonstrates that while KNN can effectively capture significant handwriting features relevant to age prediction, there remains room for improvement in addressing boundary overlaps between adjacent age groups.

8.1.2 Gender Prediction

For gender classification, KNN achieved an accuracy of 74%. The confusion matrix for gender prediction, shown in **Figure 2**, provides a detailed view of the model's performance in distinguishing between **male** and **female** handwriting. The model correctly classified 110 out of 147 female samples and 97 out of 132 male samples. Misclassifications were balanced between classes, with 37 female samples misclassified as male and 35 male samples misclassified as female. The overall performance suggests that certain handwriting traits commonly associated with gender, such as spacing and alignment, were successfully captured by the extracted features. However, the balanced misclassification rates indicate that these traits may not be fully exclusive to a particular gender.

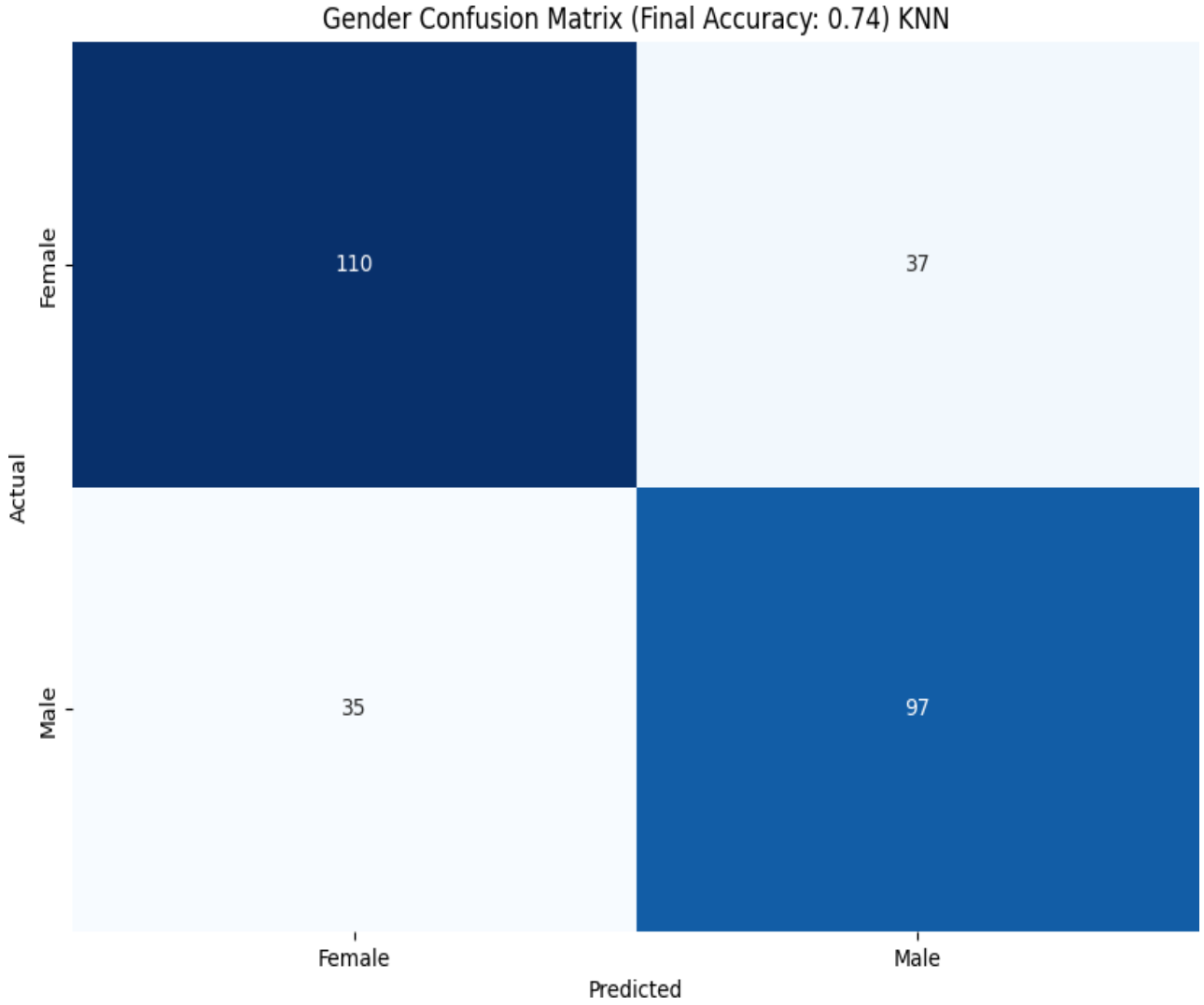

Figure 2: Confusion Matrix for Gender Classification (KNN)

The results confirm that while KNN provides a reasonable baseline for gender classification, more advanced techniques may be necessary to improve sensitivity to subtle differences in handwriting styles.

8.1.3 Handedness Prediction

KNN achieved an accuracy of 72% for handedness classification, distinguishing between **left- handed** and **right-handed** writers. The confusion matrix for handedness prediction is shown in **Figure 3**. The model correctly identified 98 out of 131 left-handed samples and 103 out of 148 right-handed samples. A notable number of misclassifications occurred, with 33 left-handed samples predicted as right-handed and 45 right-handed samples predicted as left-handed. These results suggest that the features derived from stroke orientation and curvature were partially effective in capturing handedness-related handwriting characteristics, but additional refinement in extraction may required.

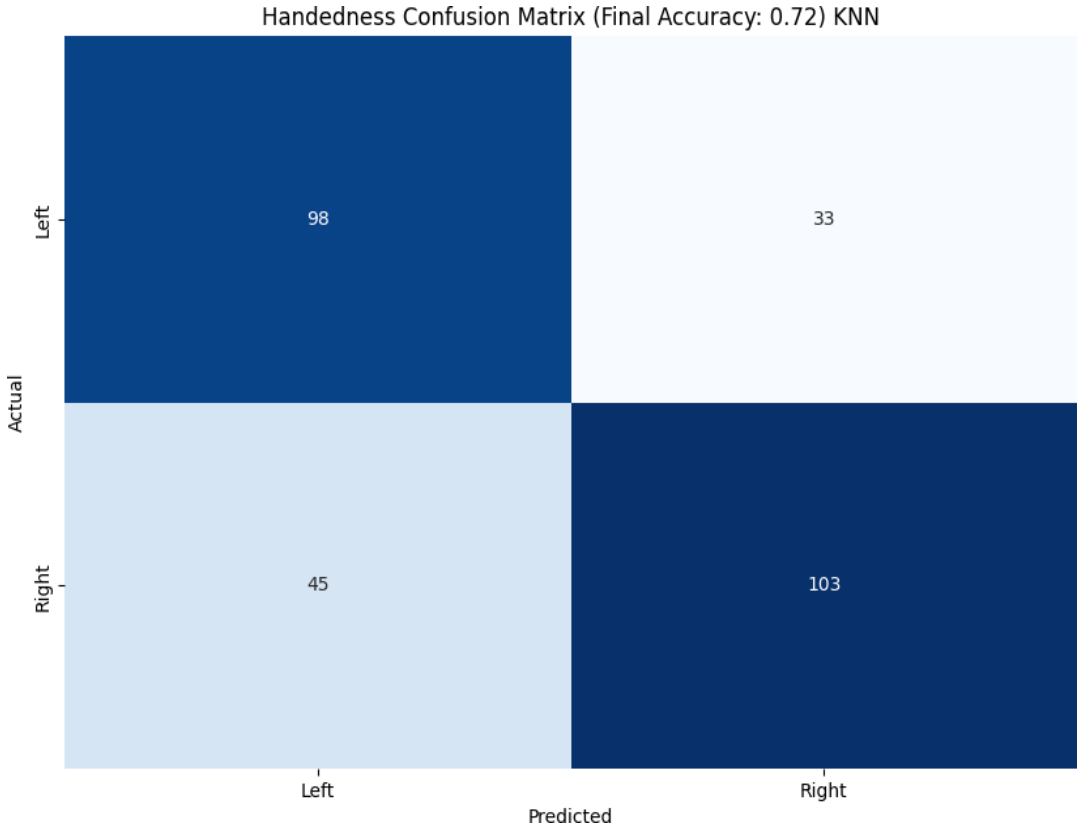

Figure 3: Confusion Matrix for Handedness Classification (KNN)

Despite the misclassifications, the overall performance highlights the utility of KNN in identifying handedness, with room for further enhancements. The overall classification accuracies for age, gender, and handedness are summarized in **Figure 4**, illustrating the comparative across the three tasks.

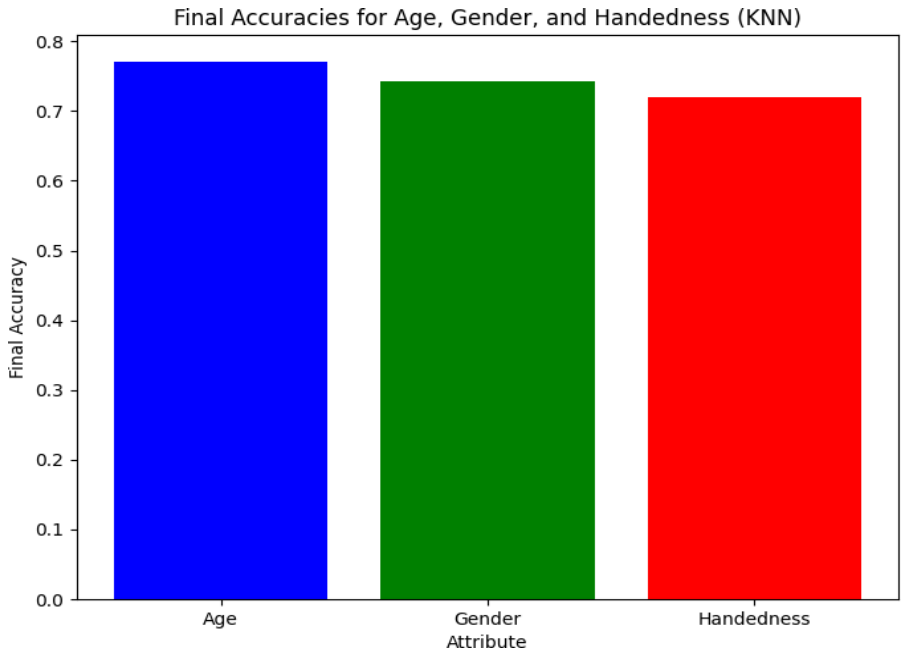

Figure 4: Final Accuracies for Age, Gender, and Handedness (KNN)

The KNN algorithm provided consistent performance across all tasks, with the highest accuracy observed for age prediction. However, the results also underscore the need for further exploration of

feature engineering and advanced classification methods to address the limitations of KNN, particularly in cases of overlapping class boundaries.

8.2 Support Vector Machine (SVM)

The performance of the Support Vector Machine (SVM) model was analyzed for its ability to classify age, gender, and handedness based on the extracted handwriting features. Utilizing a radial basis function (RBF) kernel, SVM demonstrated robust performance, leveraging its strength in capturing nonlinear decision boundaries. The results are detailed below with associated confusion matrices and accuracy comparisons.

8.2.1 Age Prediction

The SVM model achieved an accuracy of 78% for age group classification, slightly outperforming the KNN approach. The dataset was divided into the same age categories: **15–30 years**, **31–45 years**, and **46–60 years**. The confusion matrix, presented in **Figure 5**, illustrates the true versus predicted class distributions. SVM demonstrated high accuracy in classifying the **15–30 years** group, with 62 out of 96 samples correctly predicted. Misclassifications were reduced in the **46–60 years** category compared to KNN, with only 12 samples misclassified into adjacent groups. The boundary overlap between **15–30 years** and **31–45 years** groups persisted, with 34 samples from the former misclassified into the latter.

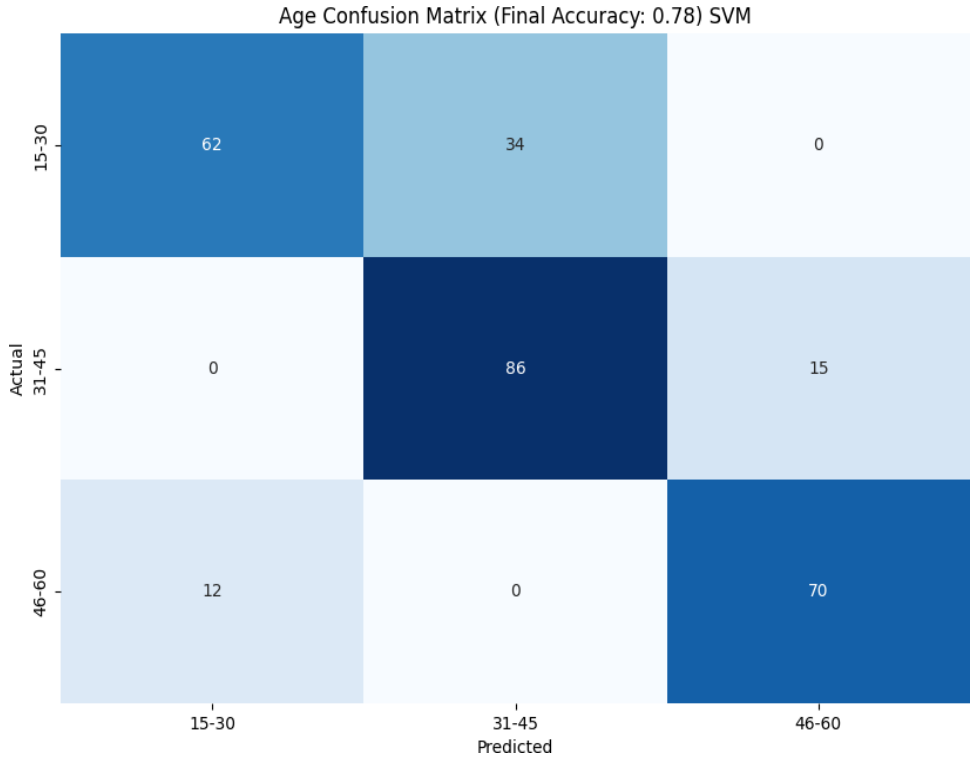

Figure 5: Confusion Matrix for Age Classification (SVM)

The performance improvement over KNN highlights SVM's capability to model complex decision boundaries, particularly in distinguishing older age groups.

8.2.2 Gender Prediction

For gender classification, the SVM model achieved an accuracy of 76%, also outperforming KNN. The confusion matrix, shown in **Figure 6**, highlights the distribution of predictions for **male** and **female** handwriting. The model correctly identified 113 out of 147 female samples and 100 out of 132 male samples. Fewer misclassifications were observed compared to KNN, with 34 female samples incorrectly

classified as male and 32 male samples classified as female. The improved performance suggests that SVM effectively captured subtle gender-specific handwriting such and alignment.

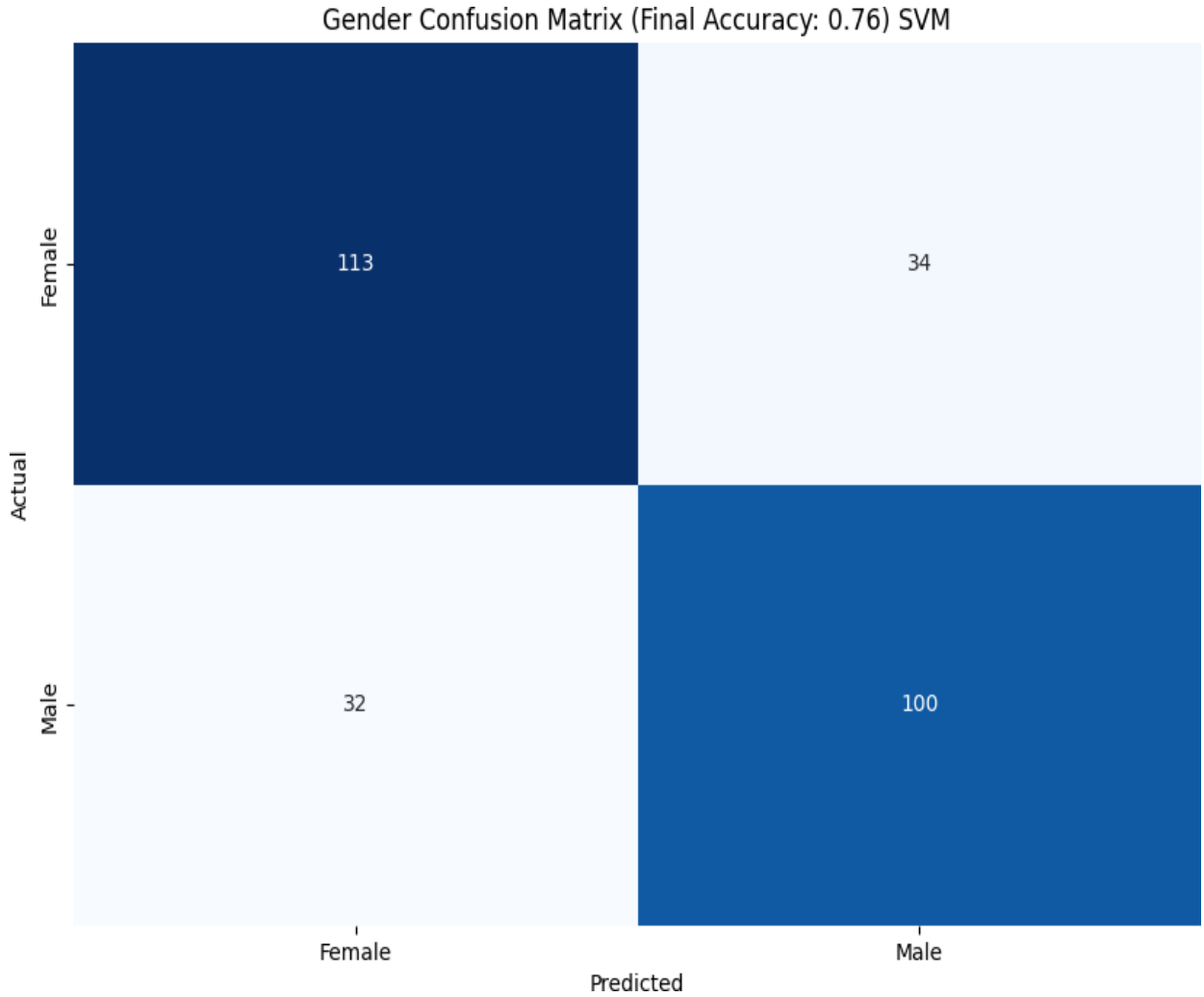

Figure 6: Confusion Matrix for Gender Classification (SVM)

The results demonstrate SVM's suitability for gender classification, reducing errors while maintaining consistent performance across classes.

8.2.3 Handedness Prediction

The SVM model achieved an accuracy of 75% for handedness classification, significantly improving on KNN's results. The confusion matrix for handedness prediction is provided in **Figure7**. The model correctly classified 103 out of 131 left-handed samples and 107 out of 148 right- handed samples. Misclassifications were reduced, with 28 left-handed samples predicted as right- handed and 41 right-handed samples predicted as left-handed. The SVM model effectively leveraged the extracted features, particularly those related to stroke orientation and curvature, to enhance classification accuracy.

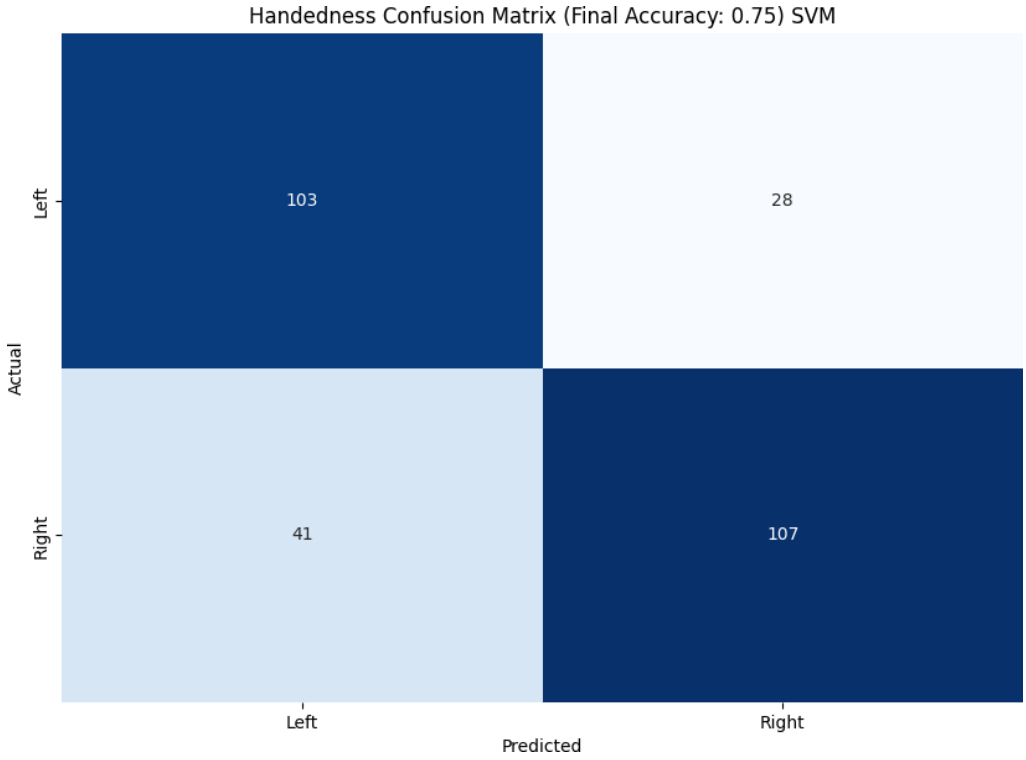

Figure 7: Confusion Matrix for Handedness Classification (SVM)

The results indicate that SVM effectively addresses challenges in handwriting-based handedness detection, particularly in reducing boundary overlap.

Accuracy Summary

The comparative accuracy results for age, gender, and handedness classification using SVM are summarized in **Figure 8**, showcasing consistent improvement over KNN across all tasks.

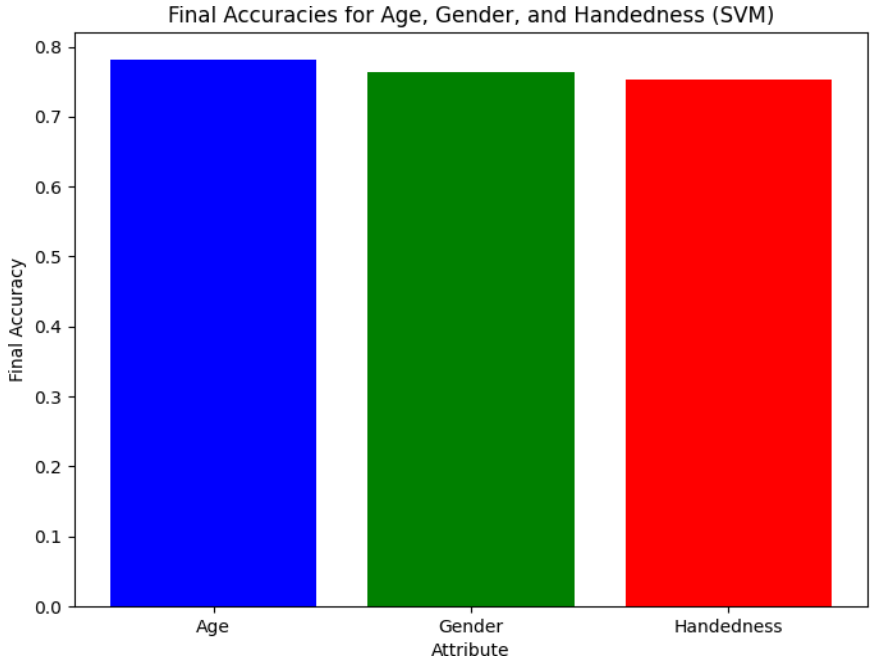

Figure 8: Final Accuracies for Age, Gender, and Handedness (SVM)

The superior performance of SVM can be attributed to its ability to model nonlinear relationships in

handwriting data, particularly in capturing fine-grained features that are critical for accurate classification. However, the results also highlight areas where further refinement is required, particularly in addressing overlapping class boundaries for age prediction. The findings establish SVM as a strong candidate for handwriting attribute classification, paving the way for evaluating the hybrid SVM-KNN model.

8.3 Hybrid SVM-KNN

The hybrid SVM-KNN model combined the global decision-making strength of SVM with the local refinement capabilities of KNN to enhance classification performance for age, gender, and handedness attributes. This section presents the results of this hybrid approach, including confusion matrices and accuracy comparisons for each classification task.

8.3.1 Age Prediction

hybrid SVM-KNN model achieved an accuracy of 79% for age group classification, outperforming both standalone SVM and KNN models. The confusion matrix for age prediction, shown in **Figure 9**, highlights the improvement in boundary separation between age groups. The model correctly predicted 64 out of 96 samples for the **15–30 years** group and 70 out of 82 samples for the **46–60 years** group. Misclassifications between adjacent groups (e.g., **31–45 years** misclassified as **46–60 years**) were slightly reduced compared to SVM, indicating improved handling of overlapping class boundaries. The **31–45 years** group maintained high accuracy, with 87 samples correctly classified.

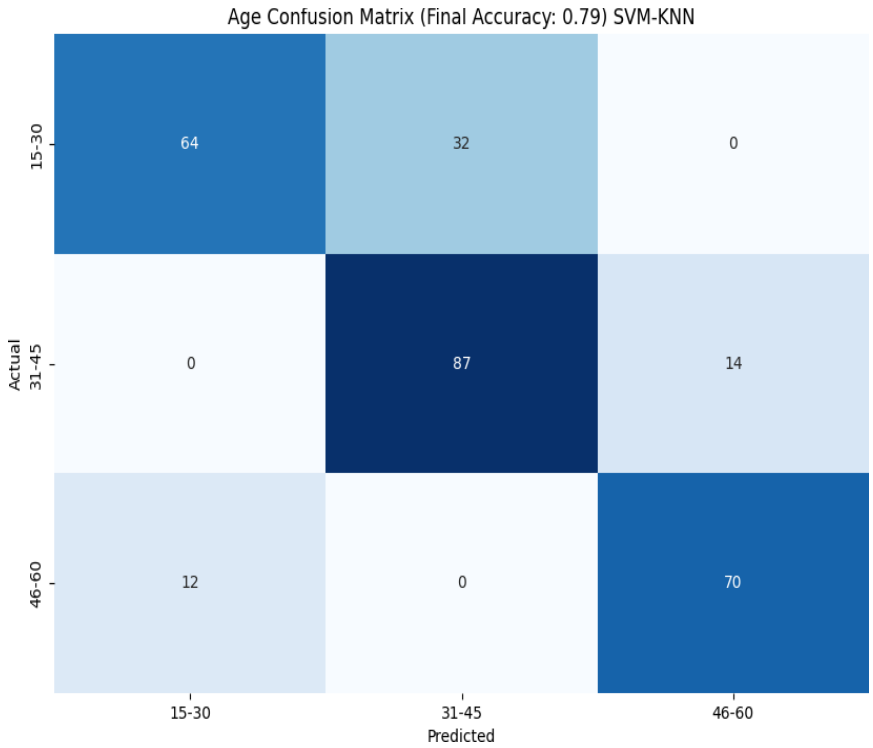

Figure 9: Confusion Matrix for Age Classification (SVM-KNN)

These results demonstrate the hybrid model's effectiveness in combining SVM's global decision boundary with KNN's local adjustment for overlapping class resolution.

8.3.2 Gender Prediction

For gender classification, the hybrid model achieved an accuracy of 80%, the highest among all models.

The confusion matrix, presented in **Figure 10**, illustrates the improvement in correctly identifying both **male** and **female** handwriting. The model correctly classified 116 out of 147 female samples and 108 out of 132 male samples. Misclassifications decreased significantly compared to standalone SVM and KNN, with only 31 female samples incorrectly predicted as male and 24 male samples predicted as female. The hybrid model leveraged SVM's robust separation capabilities while using KNN to refine boundary decisions, particularly in ambiguous cases.

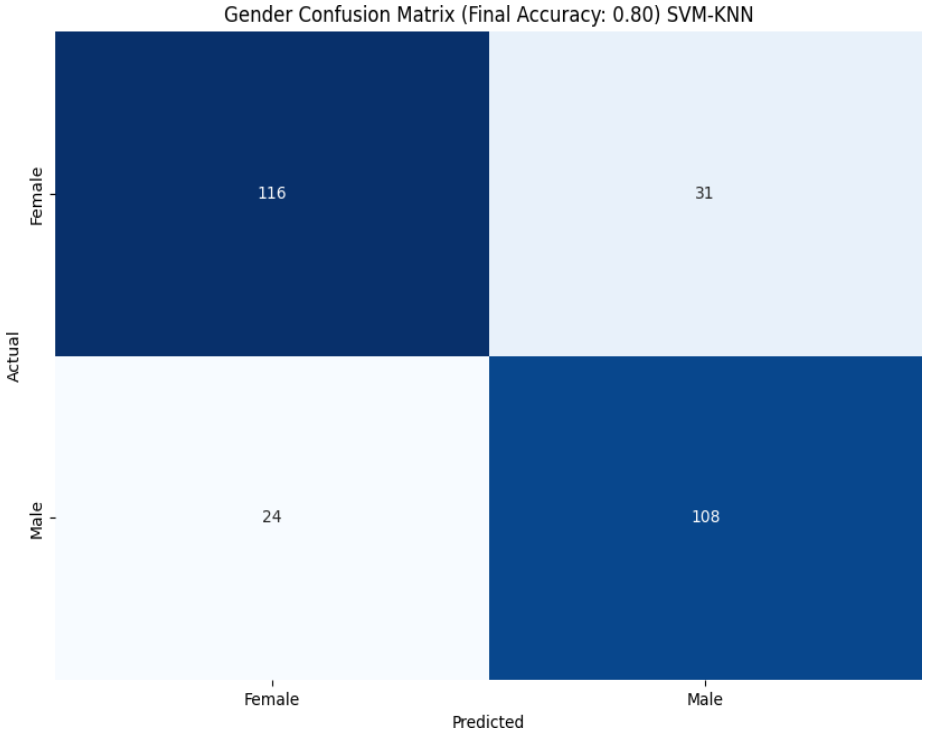

Figure 10: Confusion Matrix for Gender Classification (SVM-KNN)

The hybrid approach's superior performance underscores its ability to effectively utilize feature variations for more precise gender classification.

8.3.3 Handedness Prediction

The hybrid SVM-KNN model achieved an accuracy of 78% for handedness classification, surpassing both standalone models. The confusion matrix for handedness prediction is shown in **Figure 11**. The model correctly classified 105 out of 131 left-handed samples and 113 out of 148 right-handed samples. Misclassifications were further reduced compared to SVM, with 26 left- handed samples misclassified as right-handed and 35 right-handed samples misclassified as left- handed. The hybrid approach successfully utilized SVM's decision-making to partition the dataset globally, while KNN refined predictions within each SVM-defined class.

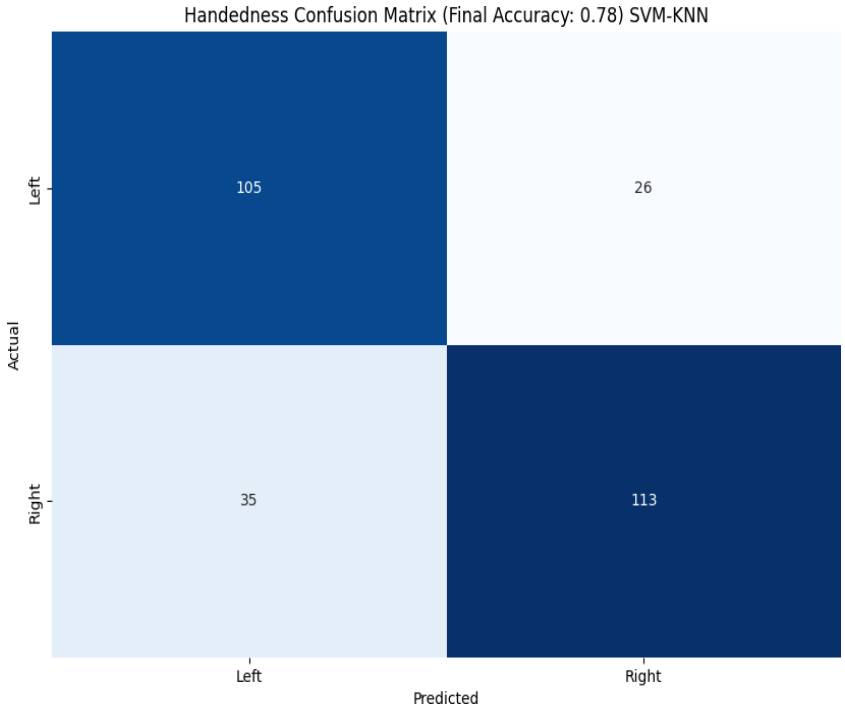

Figure 11: Confusion Matrix for Handedness Classification (SVM-KNN)

The results confirm that the hybrid model effectively balances local and global classification strengths to achieve consistent improvement in handedness detection. The comparative accuracy results for all classification tasks are summarized in **Figure 12**, illustrating the superior performance of the hybrid SVM-KNN model across all three attributes.

9. Conclusion

This study explored the use of machine learning techniques, including KNN, SVM, and a hybrid SVM-KNN model, for the classification of age, gender, and handedness from Punjabi handwriting samples. The dataset, comprising over 500 handwritten documents collected from 100+ individuals, provided a robust basis for testing the performance of these models. Feature extraction techniques such as HOG, SURF, SIFT, OGD, and wavelet transforms were used to create a comprehensive feature set that effectively captured the nuances of handwriting. The results demonstrated that while KNN offered a solid baseline, SVM outperformed it due to its ability to model nonlinear decision boundaries and better handle feature complexities. The hybrid SVM- KNN model emerged as the most effective, achieving the highest accuracy for all classification tasks by leveraging the global separation power of SVM and the local refinement capability of KNN. The hybrid approach achieved 79% accuracy for age prediction, 80% for gender prediction, and 78% for handedness classification, demonstrating its ability to resolve boundary overlaps and improve classification precision. These findings highlight the potential of combining machine learning models for complex handwriting-based attribute prediction and pave the way for future research involving larger datasets and advanced deep learning techniques to further enhance performance

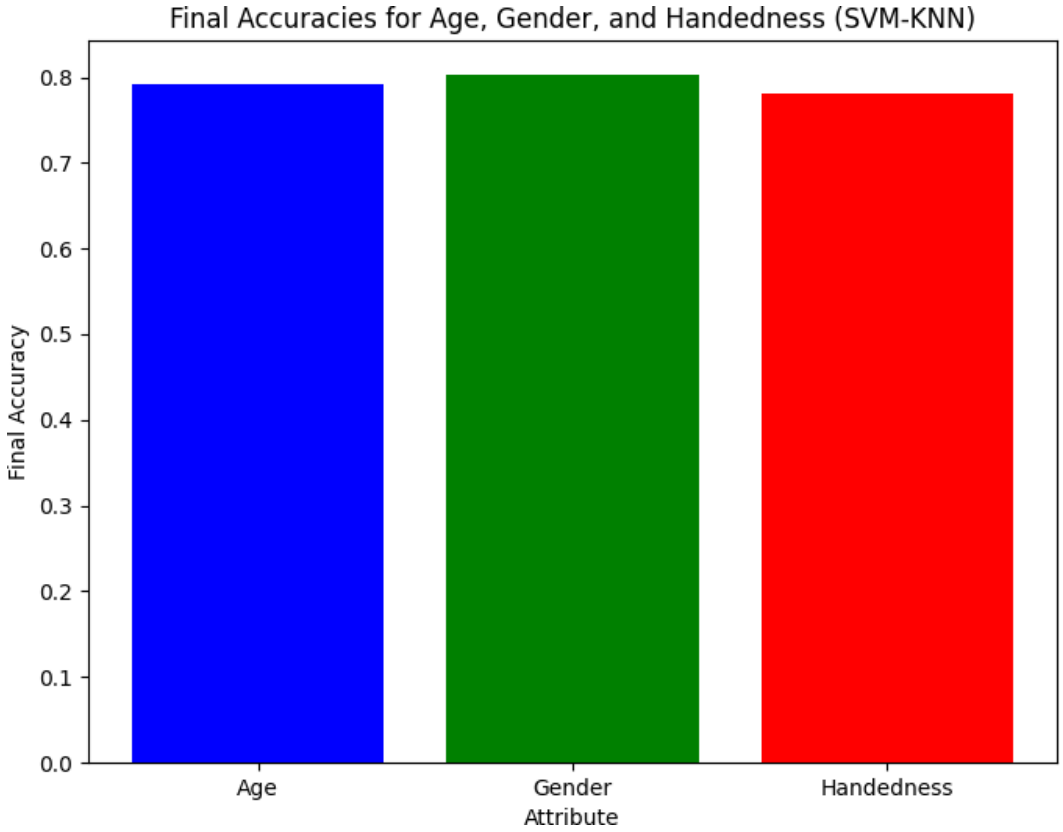

Figure 12: Final Accuracies for Age, Gender, and Handedness (SVM-KNN)

The hybrid SVM-KNN model demonstrates the best overall performance among the tested approaches, achieving the highest accuracies for all classification tasks. By combining SVM's capability to model complex decision boundaries with KNN's strength in resolving local ambiguities, the hybrid model effectively addresses challenges in handwriting-based attribute classification. These results establish the hybrid approach as the most reliable among the evaluated models.

Acknowledgments

The authors would like to express their sincere gratitude to **Dr. Manish Kumar Jindal (Supervisor)** and **Dr. Rupinderpal Kaur (Co-Supervisor)** for their invaluable guidance, continuous support, and constructive feedback throughout the development of this research work. The authors also acknowledge **Punjab University Chandigarh** for providing the necessary facilities, resources, and academic environment that enabled the successful completion of this study.

References:

1. A. Asta, E. M. Yuniarno, S. M. S. Nugroho & C. Avian, "Handwriting classification based on hand movement using ConvLSTM" *2023 International Seminar on Intelligent Technology and its Applications (ISITIA)*, 1(2023): 341-346. https://doi.org/10.1109/ISITIA59021.2023.102210
2. Gattal, C. Djeddi, A. Bensefia, & A. Ennaji , "Handwriting Based Gender Classification Using COLD and Hinge Features", *Image and Signal Processing*, 12119 (2020): 233-242. https://doi.org/10.1007/978-3-030-51935-3_25
3. Bartz, H. Yang, & C. Meinel, "Handwriting classification for art-historical document analysis", *Pattern Recognition Letters*, 138 (2020): 227-235. https://doi.org/10.1016/j.patrec.2020.06.018
4. G. Cordasco, M. Buonanno, M. Faúndez-Zanuy, M. Riviello, L. Likforman-Sulem & A. Esposito,

"Gender Identification through Handwriting: an Online Approach", *11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 1 (2020): 197-202.
5. https://doi.org/10.1109/CogInfoCom50765.2020.9237863
6. https://doi.org/10.4467/12307483pfs.22.013.17686
7. J. Dzida, "Classifying handwriting samples according to their type using discriminant analysis", *Problems of Forensic Sciences*, 94 (2023) : 1-17.
8. J. Koh'ut, M. Hradis & M. Kiss, "Towards writing style adaptation in handwriting recognition", *ArXiv*, abs/2302.06318, 1-12 (2023). https://doi.org/10.48550/arXiv.2302.06318
9. J. Long, Z. Wang, & Y. Chen, "Neural networks for handwriting recognition using optical systems. *Optics Express*, 28(22,2022): 32924-32933. https://doi.org/10.1364/OE.410950
10. J. Shin, M. A. Hasan, M. Maniruzzaman, A. Megumi, A. Suzuki, & A. Yasumura, "Online Handwriting Based Adult and Child Classification using Machine Learning Techniques", *IEEE 5th Eurasian Conference on Educational Innovation (ECEI)*, 1 (2022): 201-204. https://doi.org/10.1109/ecei53102.2022.9829467
11. J. Shin, M. Maniruzzaman, Y. Uchida, M.A. Hasan, A. Megumi, A. Suzuki & A. Yasumura, "Important Features Selection and Classification of Adult and Child from Handwriting Using Machine Learning Methods", *Applied Sciences*, 12(52, 2022): 1-12. https://doi.org/10.3390/app12105256
12. J. Y. Liu, Y. Zhang, F. Yin & C. L. Liu, "Streaming stroke classification of online handwriting", *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1(2023): 1-5. https://doi.org/10.1109/ICASSP49357.2023.10095877
13. M. Azmi, S. L. Fathima & M. Wasid, "Unveiling emotions through handwriting: A data analysis approach", 1*2th International Conference on Advanced Computing (ICoAC)*, 1 (2023): 1-6. https://doi.org/10.1109/ICoAC59537.2023.10250008
14. M. Bublin, F. Werner, A. Kerschbaumer, et al., "Handwriting evaluation using deep learning with SensoGrip", *Sensors (Basel, Switzerland)*, **23**(11, 2023): 5215. https://doi.org/10.3390/s23115215
15. M. Ghali, A. Badr & N. Ashour, "Human personality identification based on handwriting analysis using support vector machines" *Applied Artificial Intelligence*, 36(12,2022): 1021-1036. https://doi.org/10.1080/08839514.2022.2139579
16. M. Rahmanian , & M. A. Shayegan, "Handwriting-based gender and handedness classification using convolutional neural networks". *Multimedia Tools and Applications*, 80 (2021): 24573- 24602. https://doi.org/10.1007/s11042-020-10170-7
17. N. AL-Qawasmeh, & C. Suen, "Transfer Learning to Detect Age From Handwriting", *Advances in Artificial Intelligence and Machine Learning*, 2 (2022): 1-15. https://doi.org/10.54364/aaiml.2022.1126
18. P. Maken, & A. Gupta, "A method for automatic classification of gender based on text- independent handwriting", *Multimedia Tools and Applications*, 80(2021): 24573-24602. https://doi.org/10.1007/s11042-021-10837-9
19. P. Maken, & A. Gupta, "A method for automatic classification of gender based on text- independent handwriting" *Multimedia Tools and Applications*, 80 (2021): 24573-24602. https://doi.org/10.1007/s11042-021-10837-9
20. *Proceedings of SPIE*, 12787 (2023): 1278724. https://doi.org/10.1117/12.3004631
21. S. Akhtar, M. Dipti, T. A. Tinni, P. Khan, R. Kabir, & M. R. Islam, "Analysis on Handwriting Using

Pen-Tablet for Identification of Person and Handedness", *International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 1 (2021): 120-124. https://doi.org/10.1109/ICICT4SD50815.2021.9397018

22. S. Dargan & M. Kumar, "Gender Classification and Writer Identification System Based on Handwriting in Gurumukhi Script", *International Conference on Computing, Communication and Intelligent Systems (ICCCIS)*, 1(2021): 388-393, https://doi.org/10.1109/ICCCIS51004.2021.9397201
23. S. Kedar, "Identifying Learning Disability Through Digital Handwriting Analysis", *Turkish Journa of Computer and Mathematic Education*, 12 (2021): 46-56. https://doi.org/10.17762/TURCOMAT.V12I1S.1557
24. S. T. Nabi, P. Singh, & M. Kumar, "Gender Classification from Offline Handwriting Images in Urdu Script: LeNet-5 and Alex-Net", *3rd International Conference on Applied Artificial Intelligence (ICAPAI)*, 1 (2023): 1-6. https://doi.org/10.1109/ICAPAI58366.2023.10194140
25. S. Uyun, S. Rahardyan & M. Anshari, "Skew correction and image cleaning handwriting recognition using a convolutional neural network", *JOIV: International Journal on Informatics Visualization*, 7(2023): 1712. https://doi.org/10.30630/joiv.7.3.1712
26. V. Babushkin, H. Alsuradi, M. H. Jamil, M. Al-Khalil & M. Eid, "Assessing handwriting task difficulty levels through kinematic features: A deep-learning approach", *Frontiers in Robotics and AI*, 10 (2023): 1193388. https://doi.org/10.3389/frobt.2023.1193388
27. V. Ghods, "Personality recognition based on handwriting types using fuzzy inference", *IEEE Access*, 11 (2023): 86456-86469. https://doi.org/10.1109/ACCESS.2023.3303477
28. V. O, S. R & S. Shiva, "Handwritten Analysis for Gender Identification using CNN", *4th International Conference for Emerging Technology (INCET)*, 1 (2023): 1-6https://doi.org/10.1109/INCET57972.2023.10170411
29. V. Pippi, S. Cascianelli, L. Baraldi, & R. Cucchiara, "Evaluating synthetic pre-training for handwriting processing tasks" *Pattern RecognitionLetters*, 172 (2023): 44-50.https://doi.org/10.48550/arXiv.2304.01842
30. W. Cao, W. Huang, W. Guo & Y. Chen, "Handwriting removal method based on CNN",