An Enhanced Document Source Identification System for Printer Forensic Applications based on the Boosted Quantum KNN Classifier

Shahlaa Mashhadani
Department of Computer, College of Education for Pure Sciences Ibn Al-Haitham, University of Baghdad, Iraq
shahlaa.t@ihcoedu.uobaghdad.edu.iq

Wisal Hashim Abdulsalam
Department of Computer, College of Education for Pure Sciences Ibn Al-Haitham, University of Baghdad, Iraq
wisal.h@ihcoedu.uobaghdad.edu.iq (corresponding author)

Iptehaj Alhakam
Department of Computer, College of Education for Pure Sciences Ibn Al-Haitham, University of Baghdad, Iraq
ibtihaj.a.a@ihcoedu.uobaghdad.edu.iq

Oday Ali Hassen
Ministry of Education, Wasit Education Directorate, Kut, Iraq
odayali@uowasit.edu.iq

Saad M. Darwish
Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, Egypt
saad.darwish@alexu.edu.eg

Received: 26 October 2024 | Revised: 4 December 2024 | Accepted: 14 December 2024

 DOI: https://doi.org/10.48084/etasr.9420

ABSTRACT

Document source identification in printer forensics involves determining the origin of a printed document based on characteristics such as the printer model, serial number, defects, or unique printing artifacts. This process is crucial in forensic investigations, particularly in cases involving counterfeit documents or unauthorized printing. However, consistent pattern identification across various printer types remains challenging, especially when efforts are made to alter printer-generated artifacts. Machine learning models are often used in these tasks, but selecting discriminative features while minimizing noise is essential. Traditional KNN classifiers require a careful selection of distance metrics to capture relevant printing characteristics effectively. This study proposes leveraging quantum-inspired computing to improve KNN classifiers for printer source identification, offering better accuracy even with noisy or variable printing conditions. The proposed approach uses the Gray Level Co-occurrence Matrix (GLCM) for feature extraction, which is resilient to changes in rotation and scale, making it well-suited for texture analysis. Experimental results show that the quantum-inspired KNN classifier captures subtle printing artifacts, leading to improved classification accuracy despite noise and variability.

Keywords-printer forensics; document source identification; quantum-inspired computing; feature modeling

I. INTRODUCTION

Printer forensics focuses on analyzing printed documents to attribute them to a specific printer or class of printers, aiding in criminal investigations, fraud detection, and document authentication. Document source identification in this field traces a document back to its originating printer. Key challenges include: (a) Intra-printer variability, where output from the same printer changes due to factors like ink levels or paper type; (b) Inter-printer variability, as different models or units of the same model may exhibit distinct characteristics; (c) Printer aging, where wear over time alters printing characteristics, complicating source identification. Overcoming these issues is crucial for accurate forensic analysis [1-2].

Printer identification studies employ passive and active techniques [3-4]. Passive methods rely on identifying built-in features in printed documents to differentiate printer models, requiring specialized instruments and analysis tools. Active methods categorize printers based on characteristics such as banding, pattern noise, geometric distortion, profiles, and texture, which offer high accuracy even with varying fonts and printer aging. Active techniques also embed internal features, such as yellow, nearly invisible dots, encoding details such as print date and serial number, but this is mostly limited to color laser printers. Most forensic printer identification relies on passive techniques that exploit unique printer characteristics, which are difficult to alter or remove [5-7].

Printer identification for Arabic manuscripts faces challenges such as multiple character shapes, variations in paper type, font size, printer usage, and the growing complexity of document falsification through image processing tools. Some features may also be redundant, reducing accuracy and increasing processing time. Effective feature selection is critical, as it ensures that only relevant features are used to build a model that represents each printer's unique characteristics, helping to identify the source of unknown printed documents [8-9]. Feature modeling in printer forensics involves identifying and analyzing characteristics of printed documents to determine their source. Key features include geometric, print defects, texture, and microscopic attributes. Feature selection focuses on identifying the most relevant features, improving classification accuracy, and reducing computational load, which is especially important for large datasets or real-time analysis. By focusing on intrinsic features that are difficult to alter, the system becomes more resistant to manipulation or forgery [10-12].

K-Nearest Neighbors (KNN) offers several advantages in document source identification. As a lazy learner, KNN does not require a training phase, storing all data for classification during prediction, making it ideal for dynamic datasets that require frequent updates. It makes minimal assumptions about data distribution and is robust to noise, as majority voting among neighbors reduces the impact of mislabeled data. KNN easily accommodates new data without retraining and has only one key hyperparameter, the number of neighbors (k), simplifying model selection [6, 13-14]. The distance function is key in KNN, determining the similarity between data points, typically using Euclidean or Manhattan metrics. Quantum KNN (QKNN) enhances this by leveraging quantum computing to perform these calculations more efficiently. By representing data as quantum states and using quantum algorithms for distance computation and search, QKNN can achieve significant speedups, particularly with large datasets. Furthermore, quantum algorithms help mitigate the curse of dimensionality, making them more effective in large feature spaces [15-16].

A. Problem Statement

Printer identification involves determining the specific printer used to produce a document, including its make, model, or individual unit, based on physical and digital characteristics. This is essential in fields such as forensics, security, and intellectual property protection. Traditional methods, which depend on a limited set of features, can yield ambiguous results. Incorporating a wider range of features and advanced modeling improves accuracy, but high-dimensional data can cause the curse of dimensionality, reducing the effectiveness of distance metrics in KNN. For large, high-dimensional datasets, KNN distance calculations can also become computationally expensive.

B. Motivation and Contribution

QKNN offers a significant advancement in printer forensics by leveraging quantum principles to improve accuracy and efficiency. Unlike traditional KNN, which uses a fixed distance metric, QKNN explores multiple distance functions simultaneously, creating more adaptive classification models. This study focuses on a printer identification system for the Arabic letter "Waw," extracting Gray Level Co-occurrence Matrix (GLCM) features and applying bio-inspired feature selection. The QKNN classifier, combined with a "leave one out" method to calculate error rates, uses a quantum search algorithm to efficiently identify the k nearest neighbors, significantly speeding up the process.

II. STATE-OF-THE-ART APPROACHES

Document source identification research is classified by the features used (physical, mechanical, digital), the techniques applied (image processing, machine learning, AI, statistical methods), and the types of printers studied (laser, inkjet, dot matrix) [17-18]. Convolutional Neural Networks (CNNs) have been employed to automatically learn features for complex image classification, with systematic experiments showing that feature-based SVM models slightly outperform deep learning models for microscopic documents [19]. In [20], a novel method for paper identification was proposed, utilizing hybrid features by extracting texture features from images processed with GLCM and entering them into a CNN for further feature extraction. This approach achieved a 97.66% success rate in categorizing seven popular paper brands on the Korean market.

In [21], the effectiveness of ResNet as a deep neural network architecture for printed document identification was evaluated. ResNet can learn highly representative features while addressing the vanishing gradient problem. The proposed classification model was trained using multiple ResNet variants - ResNet50, ResNet101, and ResNet152 - on a large dataset of microprinted images from various printers. Mixup

augmentation was used to enhance the model's performance and generalizability, generating virtual training examples by interpolating image-label pairs. According to [17], no research has focused on text-independent identification using word images from different laser printers. To address this, laser printer types were categorized using a dataset of grayscale word images from four laser printer models, totaling 40,000 words. A combination of Local Binary Patterns (LBP) with KNN and cubic SVM classifiers was employed, along with a deep-learning CNN model. The KNN and cubic SVM classifiers achieved accuracies of 97.2% and 97.9%, respectively, while the CNN model reached 94.3%.

In [6], the aim was to identify the source printer without segmenting characters, words, or patches, using a small dataset. Three CNN-based approaches were evaluated on three distinct datasets with 1,185, 1,200, and 2,385 documents. The first method employed SVM for classification, using 13 pre-trained CNNs for feature extraction. The second method involved retraining an existing neural network for both feature extraction and classification through transfer learning. The third method used CNNs built from scratch for feature extraction, combined with SVMs for classification. The third approach achieved the best results. Recognizing printed documents in Chinese can be challenging due to the lack of unique characters, which traditional printer source identification methods rely on. To tackle this issue, in [7], a text-independent printer source identification approach was proposed, modeling the printer's timing characteristics using a graphical model. This method extracted timing features that allowed recognition without relying on specific characters. Experimental results showed the effectiveness of this strategy for document tracing. In [5] a new method was used to predict the printer model for specific documents. The dataset included 41 inkjet printer models from popular brands, where samples were collected with a limited number of printed letters under various conditions. Morphological characteristics were extracted from microscopic images. Validation results showed 98.6% accuracy when combining KNN with Quadratic Discriminant Analysis (QDA), compared to 96.3% with QDA alone. This method can enhance the forensic analysis of printed documents.

In [22], a novel method was presented to identify the source of color laser printers using a Counterfeit Protection System (CPS) pattern. At first, a Local Polar Pattern (LPP) was introduced to represent CPS patterns, ensuring scale and rotation invariance. Then, a noise-robust CPS pattern extraction technique was used, based on the LPP descriptor, and a measurement was established to assess the similarity between CPS patterns. Additionally, a mechanism for identifying source color laser printers was introduced using a one-shot learning approach to reduce computation and data storage costs. Experimental results showed that this method outperformed deep neural networks, achieving state-of-the-art performance.

In [23], a squeeze-excitation bottleneck residual network was introduced to identify the printer source of Quick Response (QR) codes. This approach leveraged the bottleneck residual block's minimal parameters and strong feature extraction capabilities while incorporating a squeeze-excitation attention module that emphasized relevant printer attributes and minimized irrelevant information, all with low computational costs. This resulted in a slight increase in parameters that significantly enhanced performance. This method achieved 98.77% identification accuracy on images captured with a smartphone, outperforming existing CNN-based approaches. In [24], the focus was on reducing the number of training features across various machine-learning models for source printer identification while maintaining high performance.

In [25], a novel method was presented to identify the source of unknown printed documents based on whether they are laser or non-laser (inkjet or photocopier), using the first application of Raman spectroscopy alongside principal component analysis and partial least squares discriminant analysis. In [26], a statistical analysis of printing patterns was performed at the microscopic level, examining the effects of printing direction, substrate, and technology. Although this study found minimal impact from printing direction, it demonstrated that shape descriptor indexes can effectively differentiate printing materials and technologies. Identification methods using SVM and random forests achieved an impressive 92% classification accuracy with complex geometric shape patterns. In [27], a printer-specific pooling descriptor was introduced that enhanced the performance of a local texture descriptor in two datasets. This pooling method excelled in cross-font scenarios using a straightforward correlation-based prediction approach, avoiding the complexity of traditional machine learning classifiers. During printing, subtle, invisible geometric distortions occur, creating unique profiles or signatures for each printer that can aid in classification. In [28], deep visual features from CNNs were examined for printer identification. Features were extracted from document images by dividing them into patches and characters, allowing text-dependent and text-independent analyses. Experiments on 20 printer documents showed patch identification at 95.52% and character identification at 98.06%. This method differs from others by using complete document images, leading to high accuracy.

In [29], an Auto-Machine Learning-based (AutoML) method was used to analyze printed documents and identify the source printer. This study compared three machine learning models and two AutoML candidates. AutoML outperformed conventional approaches, as it could accommodate varying degrees of ambiguity in printed patterns. In [30], a method for identifying the source printer and classifying documents among various printer classes was presented, using a dataset of 1,200 papers from 20 printers, including 13 laser and 7 inkjet models. This approach combined global features, such as the Histogram of Oriented Gradient (HOG), with local features from LBP descriptors. Various classifiers were used, including SVMs, decision trees, KNNs, and random forest, with the adaptive boosting classifier achieving a 96% success rate.

In [31], a method was proposed to identify color laser printers using cascaded learning of deep neural networks. The process began with training a refiner network through adversarial training on a synthetic dataset, followed by applying halftone color decomposition. The ConvNet responsible for decomposing halftone colors was then trained using the updated dataset, enhancing identification accuracy.

The printer-detecting ConvNet was trained with halftone images from candidate source printers, incorporating rotation and scaling robustness during training. According to the test results, the proposed method significantly outperformed existing approaches for color laser printer identification.

Most research in printer forensics has focused on classical machine-learning methods and physical analysis techniques to identify unique signatures and artifacts from specific printers. Although studies have shown the feasibility of identifying printer-specific signatures, challenges remain regarding accuracy, computational efficiency, and large dataset handling. Although quantum computing has potential, its application in printer forensics is still nascent. Introducing quantum approaches, particularly QKNN, could significantly advance the field. Additionally, optimizing feature extraction and quantum encoding methods tailored to printer forensics can enhance the effectiveness of quantum classifiers. By leveraging quantum computing's power and extracting optimal physical features from printed documents, more reliable and faster results can be achieved, improving precision and providing scalable solutions for large datasets.

III. METHODOLOGY

The proposed model employs the GLCM method to extract features specific to each printer, followed by a Genetic Algorithm (GA) to select a competent feature set for classification. The dataset comprises high-resolution images scanned at 1200 dpi with 8 bits per pixel (grayscale) from ten distinct printers, each with a unique model and serial number. Features are extracted from the isolated character "و" (waw), which is frequently used in Arabic as a conjunction meaning "and." Its unique shape makes it particularly suitable for printer identification due to several characteristics:

- Curved shape: The smooth and curved stroke is easily recognizable.
- Lack of diacritics: The absence of diacritical marks reduces misreading chances.
- Vertical alignment: Its vertical alignment minimizes blending with surrounding text.
- Consistency in the form: The shape remains unchanged regardless of its position in a word.
- Minimalistic design: The clean design ensures legibility even at small sizes or low resolutions, which is crucial for printer identification.

A. Image Preprocessing Stage

Preprocessing an image of the character "و" for printer identification offers several benefits, enhancing the accuracy and efficiency of the identification process. This stage included:

- Noise reduction: Remove unwanted noise from the image, ensuring that the character is accurately represented.
- Features enhancement: Enhance the character's distinct features, making it easier for the recognition system to identify it correctly.
- Standardization: Normalize the size and orientation of the character, providing a consistent input for the identification system.
- Uniform background: Eliminate variations in background that could interfere with character recognition.
- Improve legibility: Smoothing and morphological operations can improve the legibility of the character in low-quality or low-resolution images [32-33]. This study employed conversion to grayscale, binarization, noise reduction, image cropping, and thinning operations to make images of the character "و" suitable for further analysis or classification tasks.

B. Feature Extraction Based on GLCM

GLCM is a powerful tool for texture analysis, which can be highly beneficial in printer identification applications. Different printers produce unique textural patterns due to variations in printing mechanisms, toner/ink quality, and print resolution. GLCM helps in capturing these unique signatures. GLCM features are less sensitive to noise compared to pixel intensity-based features. This robustness is crucial for dealing with real-world documents that might have scanning noise or degradation. GLCM can generate multiple texture features, such as contrast, correlation, energy, homogeneity, and entropy, providing a comprehensive set of attributes for classification [34-35]. These measures can be highly discriminative for printer identification, helping to distinguish between slight variations in print textures. Calculating GLCMs at different orientations and distances captures directional and scale-dependent textural information, adding richness to the feature set. Each combination of distance and angle produces a separate GLCM. Therefore, having N distances and M angles produces $N \times M$ GLCMs [36]. In printer identification, GLCMs stand out due to their ability to capture detailed texture information and directional sensitivity, combined with robustness to noise and computational efficiency.

C. Feature Selection Using Genetic Algorithm (GA)

This study extracted 22 descriptors from a dataset of 1,000 printed documents produced by 10 different printers. Using a GA for feature selection in printer identification offers several advantages, particularly the ability to efficiently navigate complex high-dimensional search spaces to find optimal or near-optimal solutions. Given the large number of features derived from various image processing techniques, including GLCM, GAs are well-suited for exploring these high-dimensional feature spaces, as they can effectively reduce dimensionality by selecting the most relevant features, enhancing computational efficiency and model performance. By focusing on the most pertinent features, GAs improve the accuracy of the printer identification model by excluding irrelevant or redundant features, which leads to better generalization on unseen data. Therefore, with a smaller set of relevant features, the model is less prone to overfitting, offering improved performance. In this study, the fitness function for each chromosome in the population was calculated based on the QKNN classification error and the number of selected features. The main objective was to balance classification error minimization with maintaining a minimal set of descriptors.

GAs require careful configuration of several parameters to optimize solutions effectively. Key GA parameters and their typical values are as follows:

- Population size: This represents the number of potential solutions in the population. Larger populations enhance genetic diversity but increase computational cost, with typical values ranging from 20 to 1000.
- Number of generations: This is the number of iterations the algorithm runs, involving selection, crossover, and mutation. More generations allow greater exploration of the solution space but extend computation time, typically ranging from 10 to 100.
- Crossover probability: This indicates the likelihood of two individuals undergoing crossover to produce offspring. Higher probabilities promote exploration but may disrupt optimal solutions, with typical values between 0.5 and 0.9.
- Mutation probability: This is the chance that an individual will mutate, altering its genes. Higher probabilities introduce variability and can help escape local optima but might disrupt good solutions, typically ranging from 0.01 to 0.2.
- Selection method: This determines how individuals are selected for mating based on fitness. This study used roulette wheel selection, where selection probability was proportional to fitness.

D. Quantum KNN Classification Algorithm

Using a QKNN classifier for printer identification can be justified by several potential advantages that quantum computing offers over classical computing [15, 18]. Quantum computing introduces new ways to measure distances in feature space, possibly leading to more accurate identification. Quantum states can represent data in higher-dimensional spaces, which might capture more intricate details of printer characteristics [38]. This study used QKNN, where the calculation of the Euclidean distances was based on a quantum encoding with low qubit requirements and a simple quantum circuit, making the implementation particularly advantageous.

IV. RESULTS AND DISCUSSION

The prototype system for document source identification and printer recognition was implemented in a modular fashion using Google Colab with Python version 2.7. Testing was carried out on a Dell Inspiron N5110 laptop with the following specifications: 64-bit Windows 7 Home Premium, 4.00 GB RAM, and an Intel Core i5-2410M CPU running at 2.30 GHz. Table I lists the models and serial numbers of ten printers commonly used in the digital evidence laboratory. The documents were initially scanned at 1200 dpi with 8 bits per pixel, and the Arabic character "و" was extracted as a separate image set to 12 points in Times New Roman font. For printer source identification, an additional 100 images were randomly selected from the same document dataset for testing, while the training set consisted of 1,000 distinct "و" images from various printers. Accuracy was chosen as the objective metric to estimate recognition results.

TABLE I. PRINTER MODELS AND SERIAL NUMBERS

Printer	Brand	Model No.	Serial No.
P1	NashuaTec	Spc410Aficio	Q7088600729
P2	HP	Laser1102	Vnc4841849
P3	Samsung	M332x	Zdfbbjag30006mw
P4	Samsung	M332x	Zdfbbjcg300023e
P5	Hp	LaserJet1018	Cncig74912
P6	Canon	LBP3010B	MXBA909688
P7	Samsung	M332x	Zdfbbjag300008e
P8	HP	Laser1100	CED96852
P9	Canon	Sansys–Lbp 6020B	Mtma272571
P10	Ricoh	MP3350Aficio	FRHRO43547

The first set of experiments evaluated the classification accuracy for each printer using an optimal feature set of 5 to 7 features, compared to the initial 22. GA was used for feature selection to identify key features that minimize evaluation time while maintaining accuracy. Tables II and III present confusion matrices for the optimal feature set and all descriptors, respectively. These results indicate the system's effectiveness in extracting a descriptor set that accurately distinguishes texture features of printed documents. The optimal features, including contrast, similarity, mean, diagonal moment, and sum of variance from the GLCM, achieved the highest accuracy in various experiments. Additionally, an optimal feature set addresses the challenge of data point concentration in high dimensions, enhancing the significance of distance measures and simplifying the model to reduce the risk of overfitting [38].

TABLE II. CONFUSION MATRIX USING OPTIMAL DESCRIPTORS

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	10	0	0	0	0	0	0	0	0	0
P2	0	10	0	0	0	0	0	0	0	0
P3	0	0	9	1	0	0	0	0	0	0
P4	0	0	1	9	0	0	0	0	0	0
P5	0	0	0	0	9	1	0	0	0	0
P6	0	0	0	0	1	9	0	0	0	0
P7	0	0	0	0	0	0	10	0	0	0
P8	0	0	0	0	0	0	0	10	0	0
P9	0	0	0	0	0	0	0	0	10	0
P10	0	0	0	0	0	0	0	0	0	10
Accuracy	100	100	90	90	90	90	100	100	100	100

TABLE III. CONFUSION MATRIX USING ALL DESCRIPTORS

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	10	0	0	0	0	0	0	0	0	0
P2	0	10	0	0	0	0	0	0	0	0
P3	0	0	8	2	0	0	0	0	0	0
P4	0	0	1	6	3	0	0	0	0	0
P5	0	0	0	2	7	2	1	0	0	0
P6	0	0	0	0	0	8	0	0	0	0
P7	0	0	0	0	0	0	8	2	0	0
P8	0	0	1	0	0	0	1	8	0	0
P9	0	0	0	0	0	0	0	0	9	1
P10	0	0	0	0	0	0	0	0	1	9
Accuracy	100	100	80	60	70	80	80	80	90	90

The second set of experiments compared the identification accuracy of the proposed system, which utilizes GA for optimal descriptor selection and QKNN for classification, with a re-implemented printer identification system from [6] that used CNN on the same dataset. Results showed that using five

optimal features with a QKNN classifier (k=3) led to a 13% increase in accuracy compared to the same method without feature selection (22 features) and a 3% improvement over the CNN-based method (see Table III). This performance enhancement is attributed to the effective identification of printers through GA's extraction of discriminative features, aided by a multiobjective fitness function that balances recognition error with the cardinality of the selected features. Generally, the proposed feature-based classifier requires less computational power and memory than CNNs, which are computationally intensive due to their deep layers and numerous parameters.

TABLE IV. COMPARATIVE STUDY

Method	Accuracy (%)
Proposed method with optimal descriptors	96
Suggested method with all descriptors	83
Printer identification method using CNN [39]	93
Printer identification method using niching GA [40]	92

Feature selection enables faster training since the model learns from a reduced feature set rather than processing full images through multiple convolutions. Additionally, CNNs typically need large amounts of labeled data to perform well, while feature selection methods can excel in scenarios with limited data, as they are less prone to overfitting and can extract meaningful patterns from smaller datasets. Moreover, GA can adaptively identify features that are well-suited to small datasets, whereas CNNs may struggle to learn effective feature representations under such conditions [6, 12, 18]. Even with an optimal feature set, the choice of classifier significantly affects recognition accuracy. In [40], optimal printer descriptors were used with a traditional KNN classifier, which suffers from the curse of dimensionality, where the distance metrics become less informative as the dimensions increase. In contrast, the proposed model employs QKNN alongside optimal printer descriptors, effectively managing high-dimensional data through quantum algorithms, thus explaining its superior accuracy.

The relationship between classification accuracy and the number of samples per class is crucial in machine learning and pattern recognition, as the accuracy typically improves with more samples. Increased samples allow training data to better capture the variability and nuances of each class, leading to more accurate decision boundaries and reducing overfitting, helping the model generalize well to unseen data [39]. The third set of experiments examined the correlation between the identification rate of the proposed recognition system and the total number of samples per printer. The results showed that as the number of registered instances for a printer increases, the likelihood of accurate identification also increases. As expected, the recognition rate in Table V improves with additional samples, showing approximately a 3% accuracy increase for every additional 100 samples after reaching 400 cases, due to enhanced interclass diversity among printers.

TABLE V. CORRELATION BETWEEN THE IDENTIFICATION RATE OF THE PROPOSED SYSTEM AND THE TOTAL NUMBER OF SAMPLES PER PRINTER

Accuracy (%)	No. of samples
85	400
88	500
90	600
93	700
96	1000

The choice of the Arabic letter "و" (Waw) for printer identification is based on several practical and technical reasons:

- It is common in Arabic, making it likely to appear in most documents, thus a reliable candidate for analysis.
- It has a distinctive and simple shape, making it less susceptible to variations from handwriting or different fonts, allowing for consistent identification.
- As a relatively straightforward character, it reduces the complexity of recognition and segmentation tasks.
- It maintains a consistent form in all word positions (initial, medial, final, or isolated), simplifying the identification process.
- It rarely forms ligatures with other letters, making isolation and analysis easier and minimizing segmentation errors.

To confirm the suitability of "و" for printer identification in Arabic, a fourth set of experiments was conducted, with results shown in Table VI. The character "و" consists of various bends and circles that can be used to derive a distinct set of attributes for identifying each printer.

TABLE VI. RELATIONSHIP BETWEEN ACCURACY AND LETTER TYPE

Letter	Accuracy (%)
Alef "ا"	79
Sad "ص"	74
Ain "ع"	81
Waw "و"	95

The last set of experiments can empirically verify the superiority of QKNN over traditional KNN for printer identification based on both accuracy and time complexity. QKNN's potential for leveraging quantum computing resources should ideally demonstrate better accuracy and computational efficiency, especially as k increases, due to its ability to handle high-dimensional data more effectively, as shown in Table VII. For higher values of k, QKNN can evaluate a larger set of nearest neighbors based on Euclidean distances in high-dimensional spaces. This ability is crucial for tasks such as printer identification, where a broader analysis of similarities and differences across multiple features is required. The KNN-based classifier's accuracy starts lower but increases as k increases from 1 to 5. This is because a small k makes KNN very sensitive to noise in the data, leading to overfitting, while QKNN shows higher accuracy compared to KNN for the same k values. Quantum algorithms can better handle the complexities in the feature space, leading to improved accuracy

even at lower k values. The accuracy peaks around k = 5 and then decreases slightly. This is because, with higher k, KNN starts to include more distant neighbors, which may introduce irrelevant information. QKNN maintains its performance better than KNN as k increases. The quantum nature of QKNN allows it to handle the inclusion of more neighbors without a significant drop in accuracy. For KNN, the accuracy continues to decrease slightly as k becomes larger. The classifier starts to underfit as it becomes too generalized, smoothing out the decision boundaries too much. QKNN still shows a slight decline in accuracy but remains higher than KNN. Quantum processing provides a more robust mechanism for dealing with larger k values, while maintaining better performance.

TABLE VII. COMPARATIVE STUDY BETWEEN QKNN AND KNN WITH DIFFERENT *K* VALUES

***k* value**	**KNN classifier**	**QKNN classifier**
1	83	89
3	85	95
5	88	96
7	86	95
9	83	93
11	82	91
13	80	91
15	79	87
Time (for k=3)	300 ms	230 ms

The comparative results indicate that QKNN consistently outperforms KNN across different values of k. This is because QKNN leverages quantum superposition and entanglement to process multiple possibilities simultaneously, leading to more efficient and accurate neighbor identification. Classical KNN struggles with high-dimensional data due to the curse of dimensionality, where distance metrics become less meaningful. Quantum computing can handle high-dimensional spaces more efficiently, mitigating this issue. Quantum parallelism enables the simultaneous calculation of Euclidean distances between the query point and multiple data points, drastically speeding up this process compared to classical computation. Finally, the potential for real-time processing of large datasets makes QKNN suitable for applications such as real-time data analysis and decision-making [16, 41].

V. CONCLUSIONS

In forensic analysis, document verification to identify the specific printer that produced a document is crucial. Traditional methods of printer identification rely on classical algorithms and manual inspection, which can be time-consuming and less accurate when dealing with large datasets or high-dimensional feature spaces. This process involves analyzing unique characteristics left by printers, such as print artifacts, texture features, edge and contour features, and noise patterns. However, classical KNN algorithms may struggle with efficiency and scalability, particularly when processing high-dimensional data or large datasets. The novelty of this study lies in the introduction of QKNN for printer forensics, specifically aimed at enhancing the identification of printers through the analysis of the Arabic letter "و" (Waw). Key contributions include:

- Adaptive classification models: QKNN diverges from traditional KNN by employing multiple distance functions simultaneously, allowing for more flexible and adaptive classification compared to fixed distance metrics.
- Feature extraction and selection: This study utilized GLCM features and incorporated bioinspired feature selection techniques to enhance the quality and relevance of input data for classification.
- Efficient nearest neighbor identification: By integrating a quantum search algorithm within the QKNN framework, the study addresses the computational efficiency of identifying k nearest neighbors, significantly reducing processing time.
- Error rate calculation method: The implementation of a "leave one out" method for calculating error rates provides a robust evaluation framework for the proposed system's accuracy.

Overall, this study represents a significant advance in printer forensics by merging quantum computing principles with traditional classification techniques to improve both speed and accuracy in printer identification. To assess the effectiveness of QKNN compared to KNN in printer identification, a series of experiments were conducted. Classical KNN achieved an average accuracy of 85% across various k, while QKNN achieved an average accuracy of 92%, showing a noticeable improvement in correctly identifying the printer. However, there are potential limitations and challenges:

- Data quality and quantity affect QKNN classifier performance. A diverse dataset is crucial for accurate identification.
- GLCM feature effectiveness depends on various factors. Optimal feature selection is computationally expensive.
- Quantum models are powerful but less interpretable. Understanding classification decisions can be challenging.
- Model performance varies across printer models and conditions. Adapting to new models may require additional training.

Mitigating these challenges will improve printer source identification in forensic investigations. Future work includes conducting large-scale tests to evaluate the performance of QKNN with extensive datasets and in various practical scenarios, investigating other quantum algorithms and their potential applications in printer identification and other forensic tasks, and enhancing feature extraction techniques using deep learning and other advanced methods to improve the robustness and accuracy of printer identification.

REFERENCES

[1] H. Joren, O. Gupta, and D. Raviv, "Printing and scanning investigation for image counter forensics," *EURASIP Journal on Image and Video Processing*, vol. 2022, no. 1, Feb. 2022, Art. no. 2, https://doi.org/10.1186/s13640-022-00579-5.

[2] M. Kumar, S. Gupta, and N. Mohan, "A computational approach for printed document forensics using SURF and ORB features," *Soft*

Computing, vol. 24, no. 17, pp. 13197–13208, Sep. 2020, https://doi.org/10.1007/s00500-020-04733-x.

[3] R. Hamzehyan, F. Razzazi, and A. Behrad, "Printer source identification by feature modeling in the total variable printer space," *Journal of Forensic Sciences*, vol. 66, no. 6, pp. 2261–2273, 2021, https://doi.org/10.1111/1556-4029.14822.

[4] Y. F. Chen, H. H. Kao, and C. P. Yen, "An Application of Deep Learning Technology in The Recognition of Forged Documents with Color Laser Printing," *Journal of Computers*, vol. 34, no. 5, pp. 135–147, Oct. 2023, https://doi.org/10.53106/199115992023103405010.

[5] Y. Liu *et al.*, "Inkjet printer prediction under complicated printing conditions based on microscopic image features," *Science & Justice*, vol. 64, no. 3, pp. 269–278, May 2024, https://doi.org/10.1016/j.scijus.2024.03.001.

[6] N. F. E. Abady, H. H. Zayed, and M. Taha, "An Efficient Source Printer Identification Model using Convolution Neural Network (SPI-CNN)," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023, https://doi.org/10.14569/IJACSA.2023.0140386.

[7] R. Tian and Z. Zhu, "Printer Source Identification Based on Graph Model," in *2023 The 7th International Conference on Machine Learning and Soft Computing (ICMLSC)*, Chongqing China, Jan. 2023, pp. 125–131, https://doi.org/10.1145/3583788.3583806.

[8] M. J. Tsai and I. Yuadi, "Source Identification for Printed Arabic Characters," in *Proceedings of the 9th IEEE International Conference on Ubi-Media Computing*, 2016, pp. 49–53.

[9] H. M. Al-Barhamtoshy, K. M. Jambi, S. M. Abdou, and M. A. Rashwan, "Arabic Documents Information Retrieval for Printed, Handwritten, and Calligraphy Image," *IEEE Access*, vol. 9, pp. 51242–51257, 2021, https://doi.org/10.1109/ACCESS.2021.3066477.

[10] W. Abdulsalam, S. Mashhadani, S. Hussein, and A. Hashim, "Artificial Intelligence Techniques to Identify Individuals through Palm Image Recognition," *International Journal of Mathematics and Computer Science*, vol. 20, no. 1, pp. 165–171, 2024, https://doi.org/10.69793/ijmcs/01.2025/abdulsalam.

[11] H. M. Al-Dabbas, R. A. Azeez, and A. E. Ali, "Two Proposed Models for Face Recognition: Achieving High Accuracy and Speed with Artificial Intelligence," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13706–13713, Apr. 2024, https://doi.org/10.48084/etasr.7002.

[12] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Speech Emotion Recognition Using Minimum Extracted Features," in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, Fallujah, Iraq, Nov. 2018, pp. 58–61, https://doi.org/10.1109/AiCIS.2018.00023.

[13] R. H. Ali and W. H. Abdulsalam, "Attention-Deficit Hyperactivity Disorder Prediction by Artificial Intelligence Techniques," *Iraqi Journal of Science*, pp. 5281–5294, Sep. 2024, https://doi.org/10.24996/ijs.2024.65.9.39.

[14] M. Açıkkar and S. Tokgöz, "An improved KNN classifier based on a novel weighted voting function and adaptive k-value selection," *Neural Computing and Applications*, vol. 36, no. 8, pp. 4027–4045, Mar. 2024, https://doi.org/10.1007/s00521-023-09272-8.

[15] E. E. Miandoab and F. S. Gharehchopogh, "A Novel Hybrid Algorithm for Software Cost Estimation Based on Cuckoo Optimization and K-Nearest Neighbors Algorithms," *Engineering, Technology & Applied Science Research*, vol. 6, no. 3, pp. 1018–1022, Jun. 2016, https://doi.org/10.48084/etasr.701.

[16] J. Li, J. Zhang, J. Zhang, and S. Zhang, "Quantum KNN Classification With K Value Selection and Neighbor Selection," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 5, pp. 1332–1345, Feb. 2024, https://doi.org/10.1109/TCAD.2023.3345251.

[17] P. Gonasagi, S. S. Rumma, and M. Hangarge, "Text-Independent Source Identification of Printed Documents using Texture Features and CNN Model," presented at the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022), Aug. 2023, pp. 250–261, https://doi.org/10.2991/978-94-6463-196-8_20.

[18] B. Belarbi, M. E. A. Ghernaout, and T. Benabdallah, "Implementation of a New Geometrical Qualification (DQ) Method for an Open Access Fused Filament Fabrication 3D Printer," *Engineering, Technology & Applied Science Research*, vol. 9, no. 3, pp. 4182–4187, Jun. 2019, https://doi.org/10.48084/etasr.2689.

[19] M. J. Tsai and I. Yuadi, "Digital forensics of microscopic images for printed source identification," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 8729–8758, Apr. 2018, https://doi.org/10.1007/s11042-017-4771-1.

[20] J. Lee, H. Kim, S. Yook, and T. Y. Kang, "Identification of document paper using hybrid feature extraction," *Journal of Forensic Sciences*, vol. 68, no. 5, pp. 1808–1815, 2023, https://doi.org/10.1111/1556-4029.15330.

[21] D. T. Nguyen, P. Q. Nguyen, and H. B. A. Mai, "Analysis of printed document identification based on Deep Learning," *CTU Journal of Innovation and Sustainable Development*, vol. 15, no. Special issue: ISDS, pp. 119–125, Oct. 2023, https://doi.org/10.22144/ctujoisd.2023.042.

[22] Z. Li and Q. Peng, "Local Polar Pattern for Source Color Laser Printer Identification," *IEEE Access*, vol. 12, pp. 83377–83390, 2024, https://doi.org/10.1109/ACCESS.2024.3407205.

[23] Z. Guo, S. Wang, Z. Zheng, and K. Sun, "Printer source identification of quick response codes using residual attention network and smartphones," *Engineering Applications of Artificial Intelligence*, vol. 131, May 2024, Art. no. 107822, https://doi.org/10.1016/j.engappai.2023.107822.

[24] Q. P. Nguyen, N. T. Dang, A. Mai, and V. S. Nguyen, "Features Selection in Microscopic Printing Analysis for Source Printer Identification with Machine Learning," in *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications*, 2021, pp. 210–223, https://doi.org/10.1007/978-981-16-8062-5_14.

[25] M. N. M. Asri, N. F. Nestrgan, N. A. M. Nor, and R. Verma, "On the discrimination of inkjet, laser and photocopier printed documents using Raman spectroscopy and chemometrics: Application in forensic science," *Microchemical Journal*, vol. 165, Jun. 2021, Art. no. 106136, https://doi.org/10.1016/j.microc.2021.106136.

[26] Q.-T. Nguyen, A. Mai, L. Chagas, and N. Reverdy-Bruas, "Microscopic printing analysis and application for classification of source printer," *Computers & Security*, vol. 108, Sep. 2021, Art. no. 102320, https://doi.org/10.1016/j.cose.2021.102320.

[27] S. Joshi, Y. K. Gupta, and N. Khanna, "Source printer identification using printer specific pooling of letter descriptors," *Expert Systems with Applications*, vol. 192, Apr. 2022, Art. no. 116344, https://doi.org/10.1016/j.eswa.2021.116344.

[28] M. Bibi, A. Hamid, M. Moetesum, and I. Siddiqi, "Document forgery detection using source printer identification: A comparative study of text-dependent versus text-independent analysis," *Expert Systems*, vol. 39, no. 8, 2022, Art. no. e13020, https://doi.org/10.1111/exsy.13020.

[29] P. Q. Vo *et al.*, "Auto Machine Learning-Based Approach for Source Printer Identification," in *Recent Challenges in Intelligent Information and Database Systems*, Ho Chi Minh City, Vietnam, 2022, pp. 668–680, https://doi.org/10.1007/978-981-19-8234-7_52.

[30] N. F. El Abady, M. Taha, and H. H. Zayed, "Text-Independent Algorithm for Source Printer Identification Based on Ensemble Learning," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 1417–1436, 2022, https://doi.org/10.32604/cmc.2022.028044.

[31] D.-G. Kim, J.-U. Hou, and H.-K. Lee, "Learning deep features for source color laser printer identification based on cascaded learning," *Neurocomputing*, vol. 365, pp. 219–228, Nov. 2019, https://doi.org/10.1016/j.neucom.2019.07.084.

[32] P. Dehbozorgi, O. Ryabchykov, and T. Bocklitz, "A Systematic Investigation of Image Pre-Processing on Image Classification," *IEEE Access*, vol. 12, pp. 64913–64926, 2024, https://doi.org/10.1109/ACCESS.2024.3395063.

[33] M. J. Manaa, A. R. Abbas, and W. A. Shakur, "Improving the Resolution of Images Using Super-Resolution Generative Adversarial Networks," in *Artificial Intelligence, Data Science and Applications*, 2024, pp. 68–77, https://doi.org/10.1007/978-3-031-48465-0_9.

[34] G. Prasad, V. S. Gaddale, R. C. Kamath, V. J. Shekaranaik, and S. P. Pai, "A Study of Dimensionality Reduction in GLCM Feature-Based Classification of Machined Surface Images," *Arabian Journal for Science and Engineering*, vol. 49, no. 2, pp. 1531–1553, Feb. 2024, https://doi.org/10.1007/s13369-023-07854-1.

[35] J. Rout, S. K. Das, P. Mohalik, S. Mohanty, C. K. Mohanty, and S. K. Behera, "GLCM Based Feature Extraction and Medical X-ray Image Classification Using Machine Learning Techniques," in *Intelligent Systems and Machine Learning*, Hyderabad, India, 2023, pp. 52–63, https://doi.org/10.1007/978-3-031-35078-8_6.

[36] S. Joshi, S. Saxena, and N. Khanna, "Source printer identification from document images acquired using smartphone," *Journal of Information Security and Applications*, vol. 84, Aug. 2024, Art. no. 103804, https://doi.org/10.1016/j.jisa.2024.103804.

[37] E. Zardini, E. Blanzieri, and D. Pastorello, "A quantum k-nearest neighbors algorithm based on the Euclidean distance estimation," *Quantum Machine Intelligence*, vol. 6, no. 1, Apr. 2024, Art. no. 23, https://doi.org/10.1007/s42484-024-00155-2.

[38] F. Kamalov, S. Elnaffarr, A. Cherukuri, and A. Jonnalagadda, "Forward feature selection: empirical analysis," *Journal of Intelligent Systems and Internet of Things*, vol. 11, no. 1, pp. 44–54, 2024, https://doi.org/10.54216/JISIoT.110105.

[39] W. Chen and J. Li, "SIGAN-CNN: Convolutional Neural Network Based Stepwise Improving Generative Adversarial Network for Time Series Classification of Small Sample Size," *IEEE Access*, vol. 12, pp. 85499–85510, 2024, https://doi.org/10.1109/ACCESS.2024.3413948.

[40] S. M. Darwish and H. M. ELgohary, "Building an expert system for printer forensics: A new printer identification model based on niching genetic algorithm," *Expert Systems*, vol. 38, no. 2, 2021, Art. no. e12624, https://doi.org/10.1111/exsy.12624.

[41] R. Divya and J. Dinesh Peter, "Quantum Machine Learning: A comprehensive review on optimization of machine learning algorithms," in *2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS)*, Kollam, India, Nov. 2021, pp. 1–6, https://doi.org/10.1109/ICMSS53060.2021.9673630.