An Efficient Source Printer Identification Model using Convolution Neural Network (SPI-CNN)

Naglaa F. El Abady[1], Hala H. Zayed[2], Mohamed Taha[3]

Computer Science Department, Faculty of Computers and Artificial Intelligence, Benha, Egypt[1, 2, 3]
School of Information Technology and Computer Science (ITCS), Nile University, Egypt[2]

Abstract—Document forgery detection is becoming increasingly important in the current era, as forgery techniques are available to even inexperienced users. Source printer identification is a method for identifying the source printer and classifying the questioned document into one of the printer classes. According to what we know, most earlier studies segmented documents into characters, words, and patches or cropped them to obtain large datasets. In contrast, in this paper, we worked with the document as a whole and a small dataset. This paper uses three techniques dependent on CNN to find the document source printer without segmenting the document into characters, words, or patches and with small datasets. Three separate datasets of 1185, 1200, and 2385 documents are used to estimate the performance of the suggested techniques. In the first technique, 13 pre-trained CNN were tested, and they were only used for feature extraction, while SVM was used for classification. In the second technique, a pre-trained neural network is retrained using transfer learning for feature extraction and classification. In the third technique, CNN is trained from scratch and then used for feature extraction and SVM for classification. Many experiments are done in the three techniques, showing that the third technique gives the best result. This technique achieved 99.16%, 99.58%, and 98.3% accuracy for datasets 1, 2, and 3. The three techniques are compared with some previously published papers, and found that the third technique gives better results.

Keywords—Document forgery; source printer identification (SPI); convolution neural network (CNN); transfer learning (TL); support vector machine (SVM)

I. INTRODUCTION

Investigating and analyzing digital evidence to identify the details of a crime is known as digital forensics. To find, gather, and examine digital evidence, digital forensics uses a set of specialized tools and methods [1]. In the last ten years, the use of digital documents has exploded. These digital papers may include images of official contracts, bills, checks, and other documents. Maintaining a digital document to a paper copy is easier, cheaper, and more effective, but security is a challenge.

Printed documents as disputed or questioned evidence are fairly common in forensic investigations. Due to the rise in these situations and the greater usage of printers in document creation than handwritten documents, printer inspection has become a crucial requirement in questioned document analysis in recent years. Additionally, numerous printers have been involved in the widespread forgery of printed papers during the past 20 years. In these situations, it is crucial for the investigators to decide the type of printer used and to create a connection between the disputed document and the stated printer.

Personal computers, scanners, and printers can produce forged documents such as certificates, agreements, identity cards, lottery tickets, etc. Modern printers have such high resolution that it is difficult for normal persons to differentiate forged documents from real ones. Traditional approaches use chemical techniques to detect forgeries in printed documents [2],[3]. These procedures need laboratory tools and a specialist to evaluate the samples. Additionally, these methods take a long time and risk damaging the printed paper. Digital techniques, in contrast, use a reference scanner to turn printed papers into their digital equivalent. Using digital approaches for source printer identification makes distinguishing between documents printed on various printers possible. Since all the analysis is done digitally, it is quicker and more automatic.

The two basic digital techniques for detecting source printers are extrinsic (active) and intrinsic (passive). Finding extrinsic signatures, such as watermarks, digital signatures, printer serial numbers, and printing dates, is known as active research. It's so time-consuming and expensive that using it is practically impossible. On the other side, passive characterizes the printer by identifying intrinsic features. Prior research typically used the following techniques to extract statistical features from printed documents: Feature extraction techniques include the Discrete Wavelet Transform (DWT), Local Binary Pattern (LBP), Key Printer Noise Features (KPNF), Gray-Level Co-Occurrence Matrix (GLCM), Speeded Up Robust Features (SURF), Oriented FAST Rotated and BRIEF (ORB), Histogram of gradient (HOG), spatial filters, and others[4], [5], [14]–[16], [6]–[13]. The forensic classification systems adopt support vector machine (SVM), Random Forest (RF), and ensemble techniques. However, feature extraction, feature selection, and classification operations in the abovementioned approaches require much professional human participation.

Additionally, to obtain results that may be generalized, the entire procedure must be repeated multiple times using a random selection of training and testing samples. A branch of artificial intelligence is machine learning (AI) [7]. In general, machine learning attempts to recognize the structure of data and fit that data into models that are helpful and clear. Convolution Neural Network (CNN) is an artificial neural network suggested [17]. Its network structure for shared weights is comparable to simulations of genuine biological brain networks. This feature has the potential to simplify and decrease the number of parameters in the network model. Instead of using the complicated feature extraction and data

reconstruction steps used in the conventional identification approach, CNN can use the image directly as the input. Convolutional, activation, pooling, and fully connected layers make up most of a typical CNN. Activation layers are used to enable nonlinear mapping, which enhances feature maps' capacity for expression. Fully connected layers are employed as the classifier and final output layers for classification tasks [18].

Deep learning's success is strongly connected to large amounts of data. A lack of training data can seriously affect the performance of deep learning models. Transfer learning was introduced to resolve this problem. It has many advantages, including reducing training time and improving neural network performance [19]. There are two approaches to implement transfer learning: feature extraction and fine-tuning. The pre-trained network is used as any other feature extractor when extracting features. In contrast to feature extraction, a new, fully connected head is built and layered over the base architecture when fine-tuning.

Deep learning needs a large number of datasets to give more accuracy. In this research, the problem of the large dataset is solved by using transfer learning. The transfer learning CNN was used to save time in creating a model from scratch, training, and small datasets. Transfer learning CNN is divided into two types: the first used a pre-trained model such as (AlexNet [20], VGGNet-16 [21], VGGNet-19, GoogleNet [22], ResNet-18[23], ResNet-50, ResNet-101, Inceptionv3 (24], SqueezeNet[25], XceptionNet [26], DarkNet-19[27], DarkNet-53, ShuffleNet [28]) for feature extraction, and the result was classified using other machine learning classification techniques. The second used a pre-trained model but changed the last layers. An SPI-CNN technique is developed to extract deep features that are fitted into an SVM for classification to get higher performance than with an SPI-CNN model alone. Three different data sets—the first with 10 printers and 1185 documents, the second with 20 printers and 1200 documents, and the third with 30 printers and 2385 documents—were used to test all of the earlier models. The following are some of the contributions of this work:

- The techniques are tested on 30 printers, whereas all previous studies only used a maximum of 20 printers.

- Training new CNN (SPI-CNN) from scratch adapted to this application. Despite their simplicity, neural networks have proven to be extremely successful in producing good results across all datasets.

- The proposed techniques work on a whole document without segmenting it into characters, words, or patches, which speeds up processing.

- An efficient pre-processing stage that combines histogram equalization and gamma correction is implemented, significantly improving the model's performance and increasing accuracy.

The following sections make up the entire paper. Section II briefly describes the related work for classifying the source printer of a printed text document. Section III contains a description of the specifics of our proposed approach. The

efficiency of the proposed approach is investigated using detailed experiments. The proposed approach's description and outcomes have been explored in Section IV. Lastly, conclusions from this effort are presented in Section V.

II. RELATED WORK

There are several procedures for detecting document manipulation. Most of these procedures detect the source of the printer to determine the types of printers used in the printing process [16]. The problem of source printer classification has received plenty of attention in the earlier decade [29]. This section will review the most common methods for authenticating a document and confirming that a legal printer printed it.

Mikkilineni et al. [4] introduced a printer identification process that uses an SVM classifier. They studied the impact of font size, font type, paper type, and printer age. Their printer identification technique works for various font sizes, paper types, and printer ages when those variables are constant. A novel color laser printer forensic algorithm is presented by [5]. It is based on an SVM classifier and noisy texture analysis. To estimate invisible noises, two filters are used: the Wiener filter and the 2D DWT filter. The noise texture is then analyzed using the GLCM. The machine classifier is trained and tested using 384 statistical features collected from the data. The proposed method achieves 99.3%, 97.4%, and 88.7% accuracy for brand, toner, and model recognition. In [6], the authors presented a method for detecting document forgeries based on Distortion Mutation of Geometric Parameters (DMGP) with translation and rotation distortion parameters. Both Chinese and English documents can be examined using this method. It can investigate documents based on separate characters. It is strong to JPEG compression and works well with documents of low resolution. The GLCM and DWT were utilized for texture feature extraction to examine the Chinese printed source to determine the impact of different output devices [7]. The feature selection techniques are used to choose the best feature subset, and an SVM is used to determine the source model of the documents. The average experimental results achieve a 98.64% identification rate, which is 1.27% higher than the previously known approach of GLCM. Many important statistical features, such as the Spatial filters, LBP, the Wiener filter, GLCM, the Gabor filter, DWT, Haralick, and SFTA features, are calculated using image processing techniques and data exploration techniques [8]. The highest rate of identification is achieved by the LBP method. It is considered superior to other methods in its various characteristics. In [9], presented a technique for analyzing the relationship between digital printers and printed Chinese characters. An SVM-based classification and feature selection decision fusion are used. The most significant features are methodically selected from GLCM, DWT, spatial, Wiener, and Gabor filters. The identification accuracy rate of the GLCM method gets the maximum rate compared with other approaches. In [10] presents a set of characteristics for describing geometric distortions at the text-line level. Experiments on 14 printers showed that the suggested system outperforms the current state-of-the-art method based on geometric distortion. It provides substantially higher accuracy when working with a limited training size constraint. A classifier trained using one

page, one printer, one font, three different fonts, and 14 printers had an average classification accuracy of 98.85%. In [11], a proposed system utilized all printed letters simultaneously to identify the source printer from scanned images of printed documents. All printed letters, as well as local texture pattern-based features, are classified by a single classifier. The method was tested on a public dataset of 10 printers, and a new dataset of 18 printers scanned at 600 and 300 dpi resolution and produced in four different fonts. The authors of [12] identified the document source printer using a passive technique. Some feature extraction approaches have been deployed, such as Key Printer Noise Features (KPNF), Speeded Up Robust Features (SURF), Orientated FAST rotated, and BRIEF. Three classification procedures are considered for the classification job: k-NN, Random Forest, and Decision Tree. The majority vote was for these three classification techniques. Combining ORB, KPNF, and SURF with an RF classifier and adaptive boosting approach yielded the best accuracy of 95.1%. Printer identification using GLCM is presented in [13]. A feature vector is created by extracting a set of features from each character for each letter "e" in the document. Each feature vector is then classified using a 5-Nearest-Neighbor (5-NN) classifier. With training, this approach is unaffected by font type or size, although cross-font and cross-size testing yielded mixed results. A separate 5-NN classifier block for each character would be required to classify a document using all its characters, not only "e"s. The classifier becomes more complex as a result. Techniques for color and picture documents produced by inkjet printers must also be researched. A text-independent method for adequately describing source printers using deep visual Features has been applied [14]. Through transfer learning on a pre-trained CNN, the system could recognize 1200 papers from 20 different printers, including 13 laser and 7 inkjet printers. Solutions to learn discriminant-printing patterns directly from the provided data were found by Anselmo in [15]. This enabled him to reject any past beliefs about the distinctive printing artefacts of each printer. Results of the experiments demonstrated that the technique works better than its existing counterparts and is robust to noisy data. In [16], a novel technique is proposed based on SURF, Oriented Fast Rotated, and BRIEF feature descriptors. The Random Forest, Naive Bayes, k-NN, and other classifiers combinations were used for classification. The model could correctly classify the questioned papers and assign them to the relevant printer. The accuracy was 86.5% using a combination of Naive Bayes, k-NN, Random Forest classifiers, a straightforward majority voting system, and adaptive boosting techniques. A text-independent algorithm for detecting document forgeries based on source printer identification SPI is suggested by [30]. The image is divided into the top, middle, and bottom sections. The feature extraction algorithms HOG and LBP are employed. Classification approaches such as decision trees, k-NN, SVM, random forests, bagging, and boosting are considered for printer identification. The AdaBoost classifier achieves 96% classification accuracy, which is the highest.

III. PROPOSED TECHNIQUES

In this research, three distinct techniques are proposed for SPI and classifying the questioned document into one of the printer types. The proposed techniques are worked with the document as a whole and a small dataset. In the first technique, pre-trained CNN models with transfer learning for feature extraction are used. Feature maps can be extracted from any layer to train a classical classifier. It classifies the output using an SVM classifier, which means that the SoftMax layer of a CNN model is replaced with such an SVM. In the second technique, pre-trained CNN models are adjusted via transfer learning, which involves replacing the final fully connected or learnable layer of a CNN model with a new fully connected layer equal to the number of classes in the datasets. The third technique utilizes a convolutional neural network (CNN) model to resolve the SPI problem. The suggested framework (SPI-CNN) has the ability to dynamically learn and extract printer features. The SPI-CNN and a support vector machine (SVM) classifier used in this work were trained using various datasets. The datasets description, pre-processing, and details of CNN models are covered in the subsections below.

A. Datasets Description

To test the models, three different datasets are employed. The public dataset by Khanna et al. [31] consists of 20 printers (13 laser printers and 7 inkjet printers). There are a total of 60 pages distributed to each printer. All of a printer's documents are unique. Contracts, invoices, and scientific publications are the three types of documents that are included in the dataset. The second dataset includes printed documents and extracted characters from 10 printers. English and Portuguese documents are printed on each printer. The dataset is freely available on [32]. The third dataset, which comprises 30 printers, was created by combining the first and second datasets. Details of three different datasets used in training and testing are shown in Table I.

TABLE I. DETAILS OF THREE DIFFERENT DATASETS USED IN TRAINING AND TESTING

	Dataset_1 (Khanna et al., 2007)	Dataset_2 (Ferreira et al., 2015)	Dataset_3
Name	Anselmo Ferreira dataset	Khanna dataset	Hyper dataset
Number of printers	10	20	30
Number of documents	1185	1200	2385
Resolution	4922x6530	3312x4677	3312x467 and 4922x6530
Format	.tiff	.png	.tiff and .png

B. Pre-Processing

The pre-processing phase is utilized in the training and testing phases. For the pre-processing step, there are three methods: Histogram Equalization (HE), Gamma Correction, and resizing image. Histogram equalization (HE) [33], [34], helps normalize image grey-scale values and improve brightness discrimination between foreground and background images. The histogram function is written as (1)

$$H(f) = \frac{C(f) - C(f)_{min}}{(W \times H) - C(f)_{min}} \times (G_L - 1) \quad (1)$$

Where, $H(f)$ signifies the histogram function of the image, $C(f)$ identifies the cumulative function, $C(f)_{min}$ denotes the minimum non-zero value of the cumulative distribution function, $W \times H$ gives the image's number of pixels, and G_L defines the number of grey levels utilized.

Gamma correction is a nonlinear process to manage an image's overall brightness. Translating the values of the input intensity image to new values, improves the image's contrast. The Gamma is obtained by (2),

$$N_i = G(\alpha) \times (O_i)^{\frac{1}{\gamma}} \qquad (2)$$

Where, N_i stands for the new intensity value, O_i for the old intensity value, $G(\alpha)$ stands for the gray stretch parameter utilized to linearly scale the outcome on the image of [0, 255], and γ stands for the positive constant. Gamma can have any value between 0 and 1; infinite mapping is linear when it is 1. When less than 1, Gamma is weighed in terms of greater output values. The mapping is weighted toward lower output values if Gamma exceeds 1. Fig. 1 shows three Different Gamma corrections. Finally, the input image is resized to match each model's input size because each CNN model has an input size.
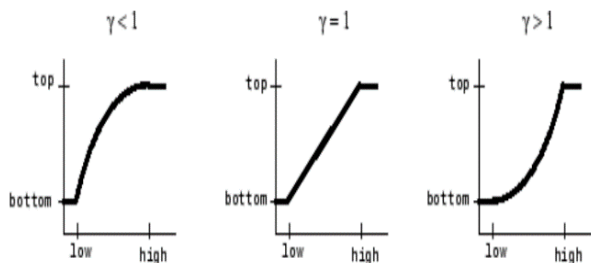

Fig. 1. Plots showing three different gamma correction settings.

C. The First Proposed Technique

Transfer learning (TL) for pre-trained models is a method that is suggested in this section for SPI. TL is used to prevent deep learning defects, which speeds up training and improves training outcomes. The pre-trained network with TL is considerably more comfortable and faster than the one trained from the start. Instead of building and training a new network, which requires millions of images, the system may quickly learn different jobs utilizing pre-trained deep networks. To transfer the learning capabilities to our application, transfer learning is employed rather than creating and training a deep learning model from scratch. 13 well-known pre-trained CNN models were used in this method to extract features (AlexNet, VGG-16, VGG-19, GoogleNet, DarkNet-19, DarkNet-53, ResNet-18, ResNet-50, ResNet101, SqueezeNet, XceptionNet, shuffleNet, inceptionv3). The obtained features are classified using SVM. In order to extract learned features from printer images, the pre-trained VGG-16 CNN model is used. The features are taken from one of the CNN layers and used to train an SVM classifier. Fig. 2 displays the model that extracts features using feature extraction and then classifies them using an SVM classifier.
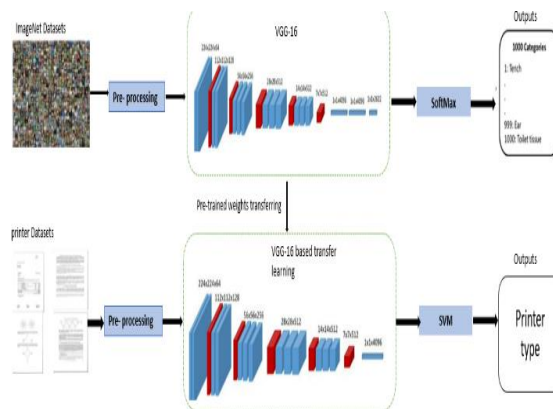

Fig. 2. Illustration of the first proposed SPI model based on transfer learning via extracting features and classifying using an SVM classifier.

D. The Second Proposed Technique

Transfer learning is based on using a CNN model that has been pre-trained and its weights that have been trained on enough data [35]. You can save time by using a pre-trained CNN model rather than creating a CNN from scratch, which requests a large, labeled dataset and lots of computational resources. While being preserved in other layers, the weights of the pre-trained CNN model are tuned in some. The higher layers of a pre-trained CNN (like DarkNet-53), initially designed for printer classification, are swapped out for the dense layer(s) in the proposed SPI approach to make the CNN compatible with SPI. In this technique, after the scanned documents for each dataset have been pre-processed. The number of classes in the current classification task is modified in every last FC layer neuron of pre-trained ConvNets ((AlexNet, VGG-16, VGG-19, GoogleNet, DarkNet-19, DarkNet-53, ResNet-18, ResNet-50, ResNet101, SqueezeNet, XceptionNet, shuffleNet, inceptionv3). With a very small learning rate of 0.0001 and 16 different batch sizes, the Adam Optimizer, also known as the Adaptive Learning Rate Algorithm, is employed to fine-tune the network. It is more efficient and less memory intensive. A model using transfer learning through fine-tuning is shown in Fig. 3.
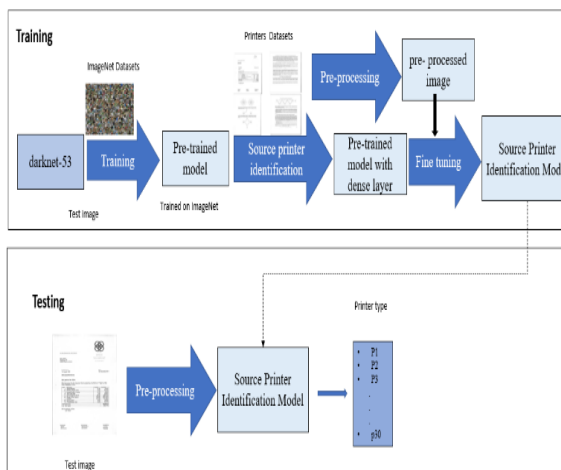

Fig. 3. Illustration of the second proposed SPI model based on transfer learning via fine-tuning in deep CNN.

E. The Third Proposed Technique (SPI_CNN)

The third technique suggests the CNN model as a solution to the SPI problem. The suggested framework (SPI-CNN) can dynamically learn and feature extraction for printers. This method uses a support vector machine (SVM) as the classifier and the SPI-CNN as the feature extraction technique.

1) SPI-CNN for features extraction: Four distinct models (7, 10, 13, and 17 layers) are used in SPI to choose the model with the highest degree of accuracy. Table II provides information about the various SPI-CNN models that were employed. Section B explains that all the dataset's scanned documents have been pre-processed. Following the pre-processing of all documents, the SPI-CNN model is applied, as shown in Fig. 4.
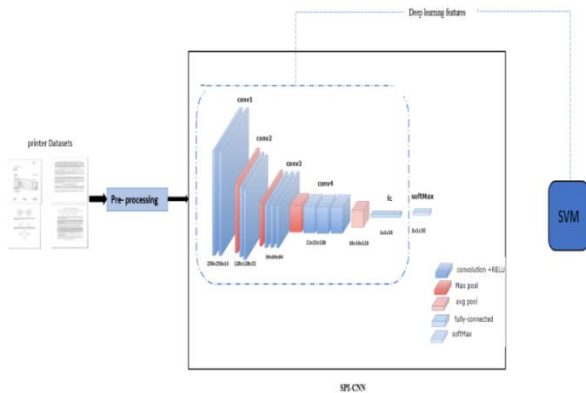

Fig. 4.   The architecture of the third proposed SPI-CNN model.

TABLE II.   DETAILS OF DIFFERENT MODELS OF CNN

Model_1	Model _2	Model _3	Model _4
Input Layer [256 256] Conv1(5x5,16) ReLU-Layer AvgPool1(2x2) Dropout Layer (0.5) Fully Connected (num class) SoftMax Layer	Input Layer [256 256] Conv1(5x5,16) ReLU-Layer AvgPool1(2x2) Conv2(5x5,32) ReLU-Layer AvgPool2(2x2) Dropout Layer (0.5) Fully Connected (num class) SoftMax Layer	Input Layer [256 256] Conv1(5x5,16) ReLU-Layer AvgPool1(2x2) Conv2(5x5,32) ReLU-Layer AvgPool2(2x2) Conv3(5x5,64) ReLU-Layer AvgPool3(2x2) Dropout Layer (0.5) Fully Connected (num class) SoftMax Layer	Input Layer [256 256] Conv1(5x5,16) ReLU-Layer AvgPool1(2x2) Conv2(5x5,32) ReLU-Layer AvgPool2(2x2) Conv3(5x5,64) ReLU-Layer AvgPool3(2x2) Conv4(5x5,128) ReLU-Layer AvgPool4(2x2) Dropout Layer (0.5) Fully-Connected (num class) SoftMax Layer

The SPI-CNN is made up of layers arranged as follows:

- A multi-layer neural network is made up of different combinations of convolutional layers with a kernel size of 5 x 5 and (16,32,64,128) number of filters.

- A 2x2 kernel size average-pooling layer is used to aggregate the generated feature maps.

- Using a dropout layer with a probability of 0.5, we generate more robust features by randomly omitting various subsets throughout training.

- The final dense layer, which used a SoftMax function and various output neurons depending on the dataset, served as the classifier.

- ReLu was utilized as the activation function in each convolutional layer to learn complex functional mappings.

2) Classification with SVM: Although SoftMax succeeds in classification, current research has shown that the SVM classifier increases classification accuracy [36]. The SVM classifier in the current investigation replaced the SoftMax layer. To train the SVM, the outputs from the layer before (FC) are used as features. After training, it applies an SPI using the features gathered from the testing image.

IV. EXPERIMENTS AND DISCUSSIONS

This section describes the experimental technique and analyses the results. All of the earlier techniques were tested on three distinct data sets: the first with 10 printers and 1185 documents, the second with 20 printers and 1200 documents, and the third with 30 printers and 2385 documents. Evaluation of the first proposed technique's performance is discussed in Section 4.1, along with the performances of the second and third proposed techniques, which are considered in Sections 4.2 and 4.3, respectively. Finally, a discussion and comparison of the three methods to other techniques are provided.

The performance of the proposed techniques is estimated using accuracy metrics [37], [38], [39]. The accuracy is obtained using the following equation (3). It is defined as the percentage of perfectly classified images, where TP: True Positive, FN: False Negative, FP: False Positive, and TN: True Negative.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

The suggested techniques were tested on a DELL PC using the following configuration implemented in MATLAB R2021b: Windows 11 64-bit, Intel(R) Core (TM) i7-11800H @ 2.30GHz, 6GHz GPU, 16GB RAM. Several tests were run to evaluate how well the suggested techniques worked.

F. Performance of the First Proposed Technique

The proposed technique is tested against 13 different pre-trained CNN models) AlexNet, VGG-16, VGG-19, GoogleNet, DarkNet-19, DarkNet-53, ResNet-18, ResNet-50, ResNet101, SqueezeNet, XceptionNet, shuffleNet, and inceptionv3 (with three distinct datasets. The number next to the model's name indicates its depth; thus, the models chosen are various, with varying depth sizes. The tests were carried out on both randomly selected 20% data as a test (i.e., 80% as the training set). When using pre-trained CNN as feature extractors and an SVM for classification, VGG-16 claims a maximum classification rate of 82.6% for dataset 1, 87.15% for dataset 2, and 86.67% for dataset 3. The accuracy of feature extraction and classification using pre-trained CNN and SVM with three different datasets is shown in Fig. 5.
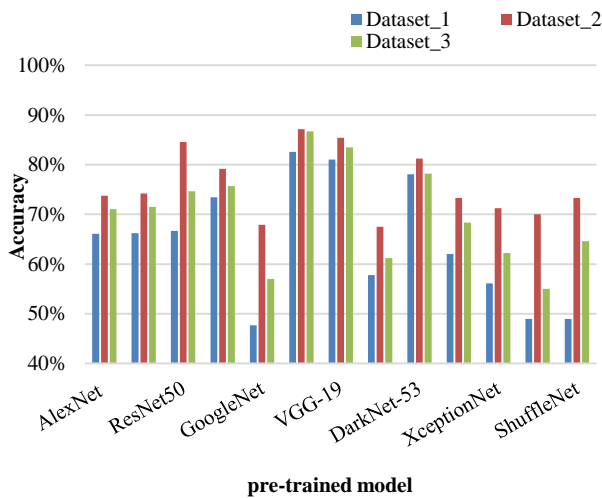

Fig. 5. The feature extraction and classification accuracy using pre-trained CNN and SVM with three different datasets (the first proposed model).

G. Performance of the Second Proposed Technique

As section D indicates, we use three separate datasets to fine-tune 11 pre-trained CNN models (AlexNet, VGG-16, VGG-19, GoogleNet, DarkNet-19, DarkNet-53, ResNet-18, ResNet-50, ResNet101, SqueezeNet, and shuffleNet). Deep transfer learning CNN architectures were used in this method to transfer learning weights, which reduced training time, mathematical calculations, and hardware resource utilization. Each dataset is split up into two parts: 20% for testing and 80% for training. The network architecture is as follows: The batch size for the 2D-CNN training was 16 samples. The Adam optimizer with a learning rate of 0.0001 was used. DarkNet-53 achieved a maximum classification rate of 98.31% for dataset 1, 97.5% for dataset 2, and 97.9% for dataset 3. Fig. 6 illustrates the performance of pre-trained CNN models during fine-tuning.
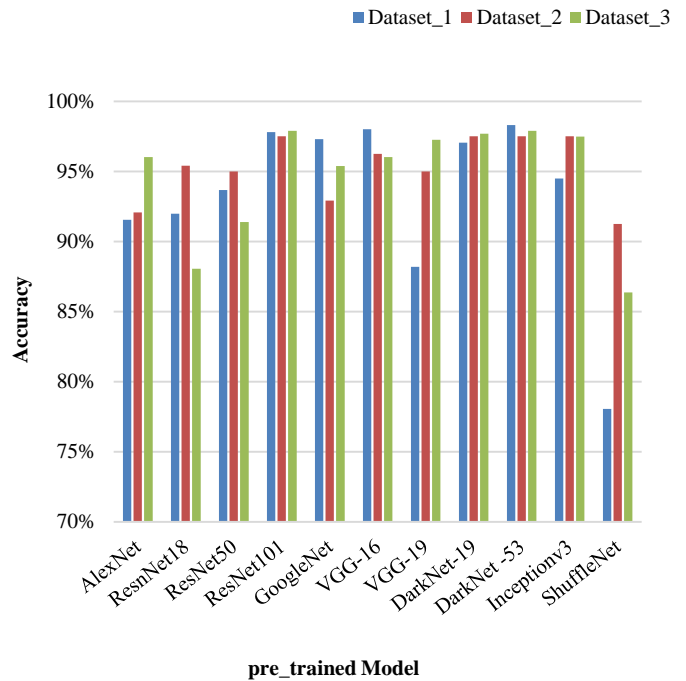

Fig. 6. The accuracy of pre-trained CNN via fine-tuning before and after pre-processing (the second proposed model).

H. Performance using the Third Proposed Technique

Three separate datasets were used to train the four CNN models that are shown in Table II. Neural network models with 1, 2, 3, and 4 convolutions were developed for comparison. Each dataset is split into 20% for testing and 80% for training. Following is the network architecture: 16 samples were used in the batch size for the 2D-CNN training. The Adam optimizer with a 0.0001 learning rate was employed. Four CNN models are used to train and classify each scanned document of a sample. Fig. 7 displays the accuracy attained by each SPI-CNN model using various datasets. Using model _4(SPI-CNN), which consists of four convolution layers, the average accuracy was 96.23% for dataset 3, 93.33% for dataset 2, and 96.2% for dataset 1. Fig. 8 displays the accuracy attained by each CNN+SVM model using a different dataset. Using model _4, which consists of four convolution layers, the average accuracy for datasets 1, 2, and 3 was 99.16%, 99.58%, and 98.32%, respectively. Results indicate that model 4 has the highest accuracy rate for SPI-CNN with SoftMax and SPI-CNN with SVM. Additionally, compared to its original configuration, the SVM classifier increased SPI-CNN accuracy by about 3%.
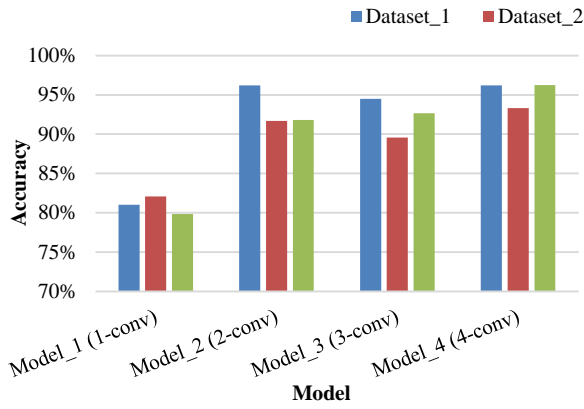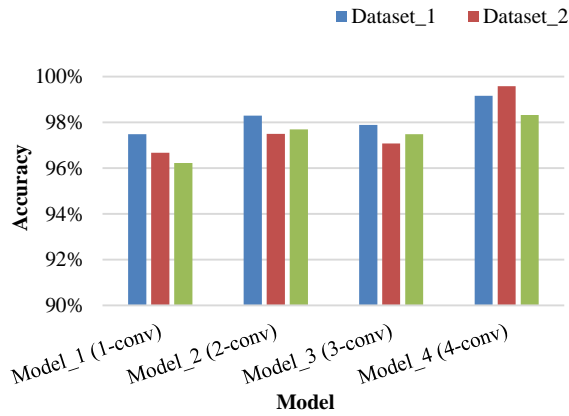
Fig. 7.   Performance of SPI_CNN model.

Fig. 8.   Performance of SPI_CNN with SVM.

Fig. 9.   Comparison of the third proposed technique with previous work on the dataset _1 of 10 printers.

Fig. 10. Comparison of the third proposed technique with previous work on the dataset _2 of 20 printers.
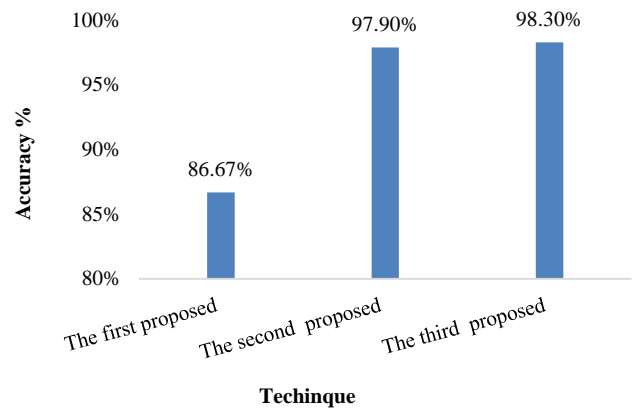
Fig. 11. Comparison of the three proposed techniques on the dataset _3 of 30 printers.

I. *Discussion and Comparison*

This section compares and discusses the results of three techniques using different datasets. Without dividing the document into letters, words, or patches and using only small datasets, three different CNN were trained to recognize the SPI. The first method was trained using SVM and used simply for feature extraction. The second had been trained in feature extraction and classification techniques. For feature extraction and classification, the third is trained completely from scratch. Because its parameters were tuned to extract features from printers document rather than other images, the third technique extracted features more effectively than the others. As illustrated in Fig. 9, our third proposed technique (SPI-CNN) outperforms [15], [40], [4], and [32] on both textural and deep learning features. As shown in Fig. 10, our third proposed technique (SPI-CNN) outperforms [37], [14], [12], [12], and [30] on dataset 2 of 20 printers and 1200 documents for both textural and deep-learned features. Fig. 11 compares the outcomes of the three proposed techniques for the data set 3 of 30 printers and 2385 documents. The prior outcomes lead us to the conclusion that the third model performs better than any previous method.
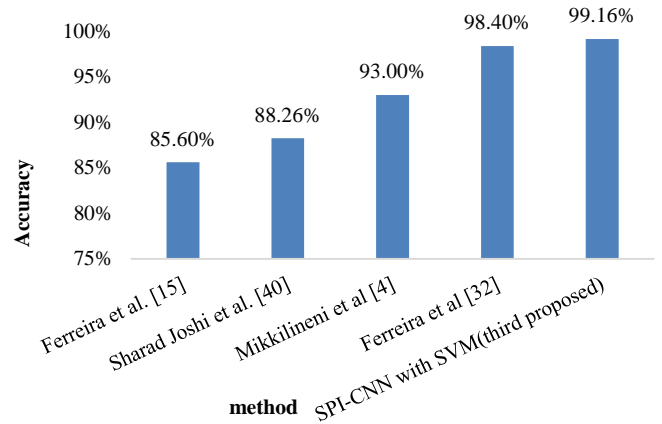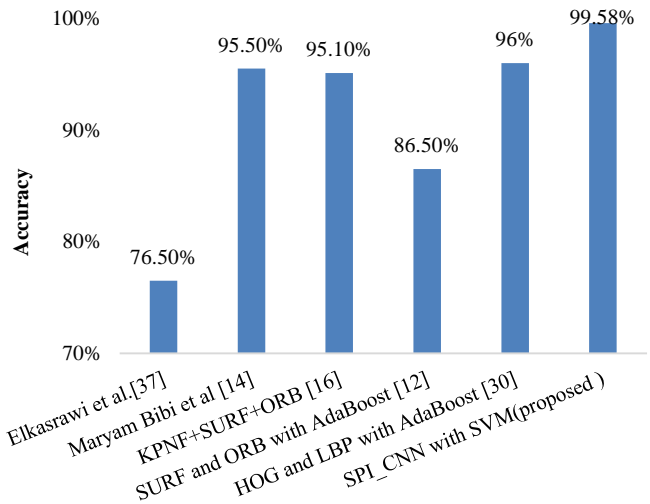
V. CONCLUSION

Three different techniques with CNN are proposed in this research to determine the printer's source. Although much research on source printer identification has been proposed, they have all been analyzed using distinct datasets and experimental setups. As earlier mentioned, several researchers use isolated characters in a text-dependent framework for experimental purposes. This paper uses CNNs to identify the source printer without segmenting the document into characters, words, or patches and with small datasets. An efficient pre-processing stage that combined histogram equalization and gamma correction was implemented, significantly improving the model's performance and increasing accuracy. The techniques are tested on a large number of 30 printers, whereas all previous studies only used a maximum of 20 printers. This paper trains three different CNN models on three separate datasets to determine the most accurate model. Transfer learning is used in the first technique for 13 pre-trained CNN models. These models serve as feature extractors, while SVM serves as a classifier. VGG-16 with SVM produces the best results. We tried 11 pre-trained models in the second technique but fine-tuned them by retraining each model and altering the last fully connected (The learning) layer. The fine-tuned DarkNet-53 achieves maximum classification rates. New CNN (SPI-CNN) from scratch adapted to this application in the third technique. The trained model was then used for feature extraction instead of SoftMax, and SVM was utilized as a classifier. Despite their simplicity, neural networks have proven to be extremely successful in producing good results across all datasets. The accuracy of the SPI-CNN model was 96.2%, 93.33%, and 96.23% for datasets 1, 2, and 3, respectively. For datasets 1, dataset 2, and dataset 3, the SPI-CNN-SVM model achieved 99.16%, 99.58%, and 98.3% accuracy, respectively. Based on the outcomes of the three techniques, we find that SPI-CNN with SVM is more accurate than the other two models. Additionally, the SVM classifier increased SPI-CNN accuracy by about 3% compared to its original configuration. With some previously published papers, the three techniques found that the third technique gives better results.

VI. FUTURE WORK

Future work will focus on discovering novel techniques to increase the accuracy of printer source identification. Try k-fold cross-validation as well rather than 20-80 validation. Identify forgeries in handwritten documents by looking at the type of ink used and the signature.

REFERENCES

[1] J. A. Lewis, "Forensic document Eeamination: fundamentals and current trends", Elsevier, Oxford. 2014.

[2] A. Braz, M. López-López, and C. García-Ruiz, "Raman spectroscopy for forensic analysis of inks in questioned documents", Forensic Sci. Int., vol. 232, no. 1–3, pp. 206–212, 2013.

[3] P. C. Chu, B. Y. Cai, Y. K. Tsoi, R. Yuen, K. S. Y. Leung, and N. H. Cheung, "Forensic analysis of laser printed ink by X-ray fluorescence and laser-excited plume fluorescence", Anal. Chem., vol. 85, no. 9, pp. 4311–4315, 2013.

[4] A. K. Mikkilineni et al., "Printer forensics using SVM techniques", Int. Conf. Digit. Print. Technol., pp. 223–226, 2005.

[5] J. H. Choi, H. Y. Lee, and H. K. Lee, "Color laser printer forensic based on noisy feature and support vector machine classifier", Multimed. Tools Appl., vol. 67, no. 2, pp. 363–382, Nov. 2013.

[6] S. Shang, N. Memon, and X. Kong, "Detecting documents forged by printing and copying", EURASIP J. Adv. Signal Process., vol. 2014, no. 1, pp. 1–13, 2014.

[7] M. J. Tsai, J. S. Yin, I. Yuadi, and J. Liu, "Digital forensics of printed source identification for Chinese characters", Multimed. Tools Appl., vol. 73, no. 3, pp. 2129–2155, Oct. 2014.

[8] M. J. Tsai and I. Yuadi, "Digital forensics of microscopic images for printed source identification", Multimed. Tools Appl., vol. 77, no. 7, pp. 8729–8758, Apr. 2018.

[9] M. J. Tsai, C. L. Hsu, J. S. Yin, and I. Yuadi, "Digital forensics for printed character source identification", Proc. - IEEE Int. Conf. Multimed. Expo, vol. 2016–August, Aug. 2016.

[10] H. Jain, S. Joshi, G. Gupta, and N. Khanna, "Passive classification of source printer using text-line-level geometric distortion signatures from scanned images of printed documents", Multimed. Tools Appl., vol. 79, no. 11–12, pp. 7377–7400, Mar. 2020.

[11] S. Joshi and N. Khanna, "Single classifier-based Passive system for pource Printer classification using local texture features", IEEE Trans. Inf. Forensics Secur., vol. 13, no. 7, pp. 1603–1614, Jul. 2018.

[12] S. Gupta and M. Kumar, "Forensic document examination system using boosting and bagging methodologies", Soft Comput., vol. 24, no. 7, pp. 5409–5426, Apr. 2020.

[13] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T. C. Chiu, J. P. Allebach, and E. J. Delp III, "Printer identification based on graylevel co-occurrence features for security and forensic applications", Secur. Steganography, Watermarking Multimed. Contents VII, vol. 5681, p. 430, Mar. 2005.

[14] M. Bibi, A. Hamid, M. Moetesum, and I. Siddiqi, "Document forgery detection using printer source identification-a text-independent approach", in 2019 International Conference on Document Analysis and Recognition Workshops, ICDARW 2019, 2019, vol. 2019–January, pp. 7–12.

[15] A. Ferreira et al., "Data-Driven Feature Characterization Techniques for Laser Printer Attribution", IEEE Trans. Inf. Forensics Secur., vol. 12, no. 8, pp. 1860–1873, 2017.

[16] M. Kumar, S. Gupta, and N. Mohan, "A computational approach for printed document forensics using SURF and ORB features", Soft Comput., vol. 24, no. 17, pp. 13197–13208, Sep. 2020.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", in Proceedings of the IEEE, 1998, vol. 86, no. 11, pp. 2278–2323.

[18] X. Yu, J. Wang, Q. Q. Hong, R. Teku, S. H. Wang, and Y. D. Zhang, "Transfer learning for medical images analyses: A survey", Neurocomputing, vol. 489, pp. 230–254, 2022.

[19] D. Larsen-Freeman, "Transfer of learning transformed", Lang. Learn., vol. 63, no. SUPPL. 1, pp. 107–129, 2013.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Adv. Neural Inf. Process. Syst., vol. 2, pp. 1097–1105, 2012.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015.

[22] C. Szegedy et al., "Going deeper with convolutions", Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07–12–June, pp. 1–9, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016–Decem, pp. 770–778.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016–Decem, pp. 2818–2826, 2016.

[25] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer

parameters and <0.5MB model size", pp. 1–13, 2016, [Online]. Available: http://arxiv.org/abs/1602.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017–Janua, pp. 1800–1807.

[27] A. Redmon, J.; Farhadi, "YOLO9000: Better, faster, stronger", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, p. 7263–7271.

[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: an extremely efficient convolutional neural network for mobile devices", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[29] P. J. Chiang et al., "Printer and scanner forensics", IEEE Signal Process. Mag., vol. 26, no. 2, pp. 72–83, 2009.

[30] N. F. El Abady, M. Taha, and H. H. Zayed, "Text-independent algorithm for source printer identification based on ensemble learning", Comput. Mater. Contin., vol. 73, no. 1, pp. 1417–1436, 2022.

[31] N. Khanna, A. K. Mikkilineni, G. T. C. Chiu, J. P. Allebach, and E. J. Delp, "Scanner identification using sensor pattern noise", Secur. Steganography, Watermarking Multimed. Contents IX, vol. 6505, p. 65051K, Feb. 2007.

[32] A. Ferreira, L. C. Navarro, G. Pinheiro, J. A. do. Santos, and A. Rocha, "Laser printer attribution: Exploring new features and beyond", Forensic Sci. Int., vol. 247, pp. 105–125, Feb. 2015.

[33] W. K. Pratt, "Digital Image Processing", vol. 19, no. 3. 1994.

[34] Anjani Suputri Devi D and Satyanarayana Ch, "An efficient facial emotion recognition system using novel deep learning neural network-regression activation classifier", Multimed. Tools Appl., vol. 80, no. 12, pp. 17543–17568, 2021.

[35] W. Tao, M. Al-Amin, H. Chen, M. C. Leu, Z. Yin, and R. Qin, "Real-time assembly operation recognition with fog computing and transfer learning for human-centered intelligent manufacturing", Procedia Manuf., vol. 48, pp. 926–931, 2020.

[36] Y. Tang, "Deep Learning using support vector machines", Accessed: Jan. 07, 2023. [Online]. Available: http://code.google.com/p/cuda-convnet.

[37] S. Elkasrawi and F. Shafait, "Printer identification using supervised learning for document forgery detection", Proc. - 11th IAPR Int. Work. Doc. Anal. Syst. DAS 2014, pp. 146–150, 2014.

[38] Y. Zhang, J. Goh, L. L. Win, and V. Thing, "Image region forgery detection: A deep learning approach", undefined, vol. 14, pp. 1–11, 2016.

[39] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-Class classification: an Overview", pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756.

[40] S. Joshi, M. Lomba, V. Goyal, and N. Khanna, "Augmented data and improved noise residual-based CNN for printer source identification", in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 2002–2006.