On the spectroscopic examination of printed documents by using a field emission scanning electron microscope with energy-dispersive X-ray spectroscopy (FE-SEM-EDS) and chemometric methods: application in forensic science

Neha Verma [1,2] · Vishal Sharma [1] (ID) · Raj Kumar [1] · R. Sharma [3] · M. C. Joshi [4] · G. R. Umapathy [5] · Sunil Ohja [5] · Sundeep Chopra [5]

Abstract

The detection of computer-generated document forgeries has always been a challenging task for forensic document examiners (FDE). With the aim to support the examination processes, Schottky field emission scanning electron microscopy with energy-dispersive X-ray spectroscopy (FE-SEM-EDS) is explored as a recent tool to analyze black toners obtained from laser printers and photocopier machines. Forty samples each from the laser printer and photocopier machines are procured and studied for morphological features, elemental profile, and multivariate analysis. The acquired SEM images and spectra are evaluated to discriminate and classify the toners having a different source of origin. Multivariate analysis is applied to develop a model of classification to successfully classify the printed documents on the basis of the similarities and differences in their composition. Hierarchical cluster analysis (HCA) discriminates the printouts in the forms of groups based on their chemical composition. The laser printer and the photocopier printed documents are grouped into 11 and eight clusters, respectively, based on their elemental composition. Cross-validation is further conducted to assess the capabilities of developed principal component analysis (PCA) and linear discriminant analysis (LDA) models for the examination of printouts from unknown origin.

Keywords FE-SEM-EDS · Multivariate analysis · Printed documents · Forensic science · Hierarchical cluster analysis

Introduction

Computer-generated and photocopied documents are usually forged by means of text insertion, text manipulations, substitution/addition of a page, or creation of a new fraudulent document altogether. In the today's era of technology, it is not uncommon for forensic document examiners (FDE) to receive questioned documents with such kind of alterations for the examination and identification of the source, authorization, authenticity, and integrity purposes. These questioned documents may vary from general documents like school or university mark sheets, medical certificates, identity proofs, lease documents, wills, and property agreements to documents of interests for homeland security like counterfeit currency and forged passports and visas [1].

The availability of a large number of printers, photocopiers, and scanning devices in today's market allows the fraudsters to create duplicate documents carrying identical features with almost no defects to be caught by the naked

✉ Vishal Sharma
  vsharma@pu.ac.in

1 Institute of Forensic Science and Criminology, Panjab University, Chandigarh 160014, India

2 DFSS Fellow, Central Forensic Science Laboratory (Document Division), Chandigarh 160036, India

3 DFSS, Central Forensic Science Laboratory (Document Division), Bhopal 462003, India

4 DFSS, GEQD, Central Forensic Science Laboratory, Shimla 171003, India

5 Inter University Accelerator Centre, Aruna Asaf Ali Marg, New Delhi 110 067, India

eye making the examination of questioned documents more challenging and forgeries harder to detect. The printing market is basically led by inkjet, laser printers, and multifunction devices, where inkjet printers use liquid inks while laser printers and photocopiers use dry toner powders to perform the task of printing. However, toner-based printing has dominated the market due to its cost-effectiveness, ease of use, fast speed, and good-quality printing. The toner ingredients are put together to form complex powders that remain well-matched with the individual cartridge, printer, image development process, charge transfer, and fusing method [2]. The various laser printers and photocopier machines irrespective of their make and model share a common principle of electrophotography with very minute differences. The use of a distinctive combination of the components in printing inks/toners is usually done with the aim of generating possibilities to distinguish formulations among same and different sources. Also, the continuous changes in the formulation are done to enhance the properties of toner to obtain high-resolution images with better quality and less edge roughness. Therefore, the authenticity of the questioned document may be verified by analyzing it with physical (microscopy) or chemical methods to know the probable source of printing [3–5].

Toner powders are basically a complex mixture of particles of 8 to 10 μm in size composed of polymers/resins to bind the pigments to the paper with thermal fusing with presence of dye or pigment to provide color, charge control agents are added to manage the charge characteristics, surface additives impart good flow properties to toners, metal nanoparticles or surfactants act as dispersing agent, and wax prevents the adhesion of toner to the rollers during fusing [6, 7]. The presence of a variety of elements in toner powders as driers, charge control agents, additives, pigments, and dyes is done to provide definite properties to the toner related to dryness, flexibility, gloss, and color [8].

The manufacturing of these toners can be achieved either by conventional pulverization or by more recent chemical methods [9]. The resulting properties of the toner vary with the uniformity in the constituents, particle size and particle shape distribution, and method of preparation [10]. Because of these reasons, toner powders from different origins have different physical and chemical properties. Over time, attempts have been made to study the composition of toner powders by different chromatographic and spectroscopic methods to identify the source of questioned documents. Researchers have demonstrated the utility of methods like pyrolysis gas chromatography (Py-GC) [11, 12], UV-Vis spectroscopy (UV-Vis) [13], infrared spectroscopy (IR) [14–19], Raman spectroscopy [20, 21], and X-ray fluorescence (XRF) [18, 22] in the analysis of composition of toners in the past.

However, recent studies have shown the capabilities of direct analysis in real-time mass spectroscopy (DART-MS) [23, 24], scanning electron microscopy-energy-dispersive X-ray spectroscopy (SEM-EDS) [17, 25–29], laser ablation inductively coupled plasma mass spectroscopy (LA-ICP-MS) [29–32], and laser-induced breakdown spectroscopy (LIBS) [32–34] in providing comprehensive and rapid analysis for inks/toner chemical profile with little to no damage of the questioned document. The studies have evaluated the performance of elemental detection in gaining the evidential value for the discrimination of printed questioned documents.

In earlier works, Trzcinska utilized various analytical methods to evaluate their discrimination potential for identification of the source of the questioned document. The element profile of toner was studied with SEM-EDX along with polymeric composition with FTIR. The peak integral ratio for common elements (C, O, S, Si, P, Al) was calculated and differences among these for various toner samples were recorded to achieve significant differentiation for forensic purposes [17]. Similar studies were also conducted by some researchers [26, 27], where the utility of SEM-EDX in the characterization of toners was compared with that of FTIR.

In another study, the author successfully classified black toner samples with results obtained from XRF and FTIR. The combined information obtained could discriminate 95.8% of the total sample pairs, whereas only 90.8% sample pairs could be discriminated utilizing only the XRF data. The elemental profile showed the presence of iron (Fe) as the dominant element followed by the presence of sulfur in almost all samples [18].

The demonstration of a two-tier method consisting of X-ray fluorescence (XRF) and laser-excited plume fluorescence (PLEAF) for multi-element analysis toners was performed by Po-Chun Chu et al. The technique PLEAF was able to generate 3D elemental mapping in the toner samples along with the sequence of printing in case of overprints unlike that of XRF. Statistical tools like k-means and principal clusters were also applied to obtain correct identifications [22].

Trejos et al. [24] evaluated the performance of five major analytical methods, i.e., SEM-EDS, LA-ICP-MS, DART-MS, FTIR, and Py-GC-MS, to determine the polymeric as well as elemental differences in the toners of different origins. The study presented the usefulness, limitations, and error rates of the techniques with their potential of discriminating samples on the basis of their origin. Furthermore, the classification and comparison capabilities of PLS-DA and KNN algorithms were tested to perform the search assess from the database by calculating the magnitude of the similarities between the test sample and the existing samples in the database.

Egan and his colleagues explored the usefulness of infrared spectroscopy to analyze toners by creating a searchable spectral library. It allowed the samples from the same

sources to be matched with high accuracy. The results were then analyzed with multivariate analysis to discriminate samples into distinct groups. 95.81% of total samples were correctly classified with the help of linear discriminant analysis (LDA). Furthermore, the study was expanded to detect the presence of elements in toner samples with the help of SEM-EDS and Py-GC-MS and applying cluster analysis and principal component analysis (PCA) on the dataset. The results confirmed that the use of R-A IR, SEM, and Py-GC/MS may help forensic document examiners to obtain a substantial amount of information regarding the probable origin of the questioned document [25].

In another study, Trejos et al. conducted a series of characterization on black toners utilizing laser-based methods and SEM-EDX to compare their discrimination capability. The tests were further evaluated with statistical methods like analysis of variance (ANOVA) with Tukey's post hoc test and PCA. The results showed that although SEM-EDS has the dual advantage of providing image characterization along with elemental analysis, its utility for discrimination within various sources remains inadequate. However, both the laser-based methods (LIBS and LA-ICP-MS) gave an improvement in the results by producing 89% and 100% discrimination respectively among toner samples of different sources [28].

Furthermore, the research group used SEM-EDS and LA-ICP-MS to determine the correct association rate among different toner samples and the individual techniques' potential to discriminate samples from a same or different source with false inclusion and exclusion rates. The study showed that the results obtained by SEM-EDS towards discrimination of toners were complementary to those acquired by LA-ICP-MS. SEM-EDS was shown to have many advantages like its dual capability of providing discrimination on the basis of particle morphology as well as with elemental composition. The elements are collected only from the layer of toner present on the surface without any interference from the paper unlike that of LA-ICP-MS. Moreover, the elements with a polyatomic interface, for example, $^{16}O_2{}^+$, $^{14}N^{18}O^+$, $^{14}N^{16}O^1H_2{}^+$, $^{15}N^{18}O^1H^+$, and $^{16}O^{18}O^+$, were analyzed more efficiently with SEM-EDS [29].

Szynkowska et al. studied the isotopic composition of toners for discrimination and characterization by utilizing inductively coupled plasma time-of-flight mass spectrometry (LA-ICP-TOF-MS). It was shown that the data acquired can be further analyzed with chemometrics and a minimum set of elements that have the most discriminating power can form the basic identifying source of the toner. Moreover, the mass spectrum for the colored toners included in the study enabled clear isotopic distinction among toners of different origins [30].

More recently, LIBS and LA-ICP-MS have shown to possess great advantages to forensic scientist, i.e., improved detection limits, the speed of analysis, ease of operation, high accuracy with minimal damage to the sample, and superior sensitivity and specificity [31]. Subedi et al. examined toner samples by utilizing a tandem LIBS/LA-ICP-MS approach where the combination includes rapid screening and confirmation of the elements for the characterization of toner samples originating from different manufacturing sources. The author concluded that the tandem mode minimizes the limitations of the individual method and provides a more comprehensive and illustrative chemical characterization. Furthermore, the combined method of analysis generates accurate results in less time with an injection of small amount of sample as compared with the two separate tests, which is one of the basic requirements in real-world forensic cases [32].

Lennard and his group successfully assessed the variation in the elemental composition across the toner samples of different origins with the help of LIBS and LA-ICP-MS, by selecting the peak ratio for the element emission lines for the purpose of discrimination. The principal component analysis accounted for 99.5% of the variation in the data for the toner analysis along with 97.4% and 98.4% discrimination for 3-sigma criterion and ANOVA with Tukey's post hoc test respectively when compared with LA-ICP-MS. The author suggested that the LIBS method provides good discrimination powers for toner samples and can be adopted as a routine method for the examination of questioned documents [33].

Metzinger et al. investigated the potential of single-shot laser-induced breakdown spectra (LIBS) using statistical evaluation including linear correlation and the sum of squared deviations and overlapping integral along with multivariate curve resolution-alternating least squares (MCR-ALS) combined with classification tree and discriminant analysis (DA). It was shown that the MCR-ALS/DA method gave 83.3% accurate results for classification of printing source utilizing spectrum obtained by a single shot of laser. However, the best discrimination results were obtained in the UV range of the instrument; the authors suggested that the same can be enhanced by increasing the number of laser shots while minimizing the contribution of paper during analysis [34].

The goal of the current study is to evaluate the performance of Schottky field emission scanning electron microscopy with energy-dispersive X-ray spectroscopy (FE-SEM-EDS [Jeol JSM-7610 F]) in discrimination and classification of black toners obtained in the form of printouts from laser printers and photocopiers of various origins on the basis of topographical properties (particle shape, size, and distribution) and constituent elemental profile. The

proposed method could help FDE to generate both qualitative and quantitative results simultaneously in a most simplified and accurate way. The expediency of using FE-SEM over other methods is its quasi-nondestructive nature with the facility of an in-lens Schottky field emission electron gun which delivers a probe current of ten times than that of the conventional field emission electron gun (FEG). The combination of in-lens with a low-aberration condenser lens (ACL) enables the efficient collection of the electrons which further improves the resolution of the image. The gentle beam (GB) mode in FE-SEM allows the reduction in the landing voltage of the electrons just before they strike the specimen. This is necessary to knock out electrons only from the layer of the toner (which is usually 20 to 95 μm) to avoid interference of elements from the paper. Also, the effects of heat which may cause undesirable changes to the nature of toners are also reduced with the deceleration of electrons. All these features make Schottky FE-SEM an attractive tool for the examination of computer-generated questioned documents. Furthermore, this study is expanded to classify the samples by using multivariate functions to achieve significant conclusions. The study design can also be applied to exhibits from other fields of forensic science to achieve notable discrimination and classification.

Materials and methods

Sample collection

The print sample sets are collected from 40 different sources each for laser printer and photocopier machines on white A4 size papers (CEDAR manufacturer) with 100 grams per square meter (GSM) weight. Five replicate samples from each source are printed with uniform settings like resolution, orientation, and mode of printing throughout the study. The printouts are obtained in the form of lines and text and no appreciable differences are seen. Gloves are worn at the time sample collection and all samples are stored in a sealed envelope at the same environmental conditions to minimize the extent of foreign contamination. The details about the printer/photocopier brand, model, cartridge type, usage, and date of collection are noted carefully by the authors. The samples printed from laser printers (L) are numbered from L1 to L40 and those of photocopier machines (P) are numbered from P1 to P40. Table S1 in the Electronic Supplementary Material (ESM) shows the printed samples and their source of origin included in the present study.

Sample preparation

The sample preparation for the analysis of both printed document toner samples and pressed toner pellets is minimum and simple.

Printed document samples

Each printed document sample has five replicates and each replicate is cut into five individual segments. These printed segments are fixed on the top surface of the stub with the help of carbon tape. Usually, when electrons (negatively charged species) hit the surface of the samples, some of them get reflected, some produce secondary electrons, and some get absorbed. The absorbed electrons then interfere with the trajectories of incoming electrons, causing blurring of the image. In order to avoid these situations, samples are generally coated with platinum, gold, or platinum-gold alloy to inhibit charging, improve imaging, and reduce thermal damaging of the samples. It is noteworthy that the presence of carbon is expected in most of the toner samples; therefore, coating samples with carbon is not preferred in the present study. Thus, the samples are exposed to platinum for 60 s to form 2-nm-thick layer of platinum on the top surface before analysis. Ten blind samples from unknown source of origin to the author are also prepared in a similar manner for the purpose of cross-validation studies.

Pressed toner pellets

Dry toner powders are extracted from cartridge bins of few printers and photocopiers to make toner pellets to perform the comparative analysis with that of the printed documents. It is done to ascertain if any change in the chemical composition/morphology of the toner occurs due to thermal fusion of toners to the surface of paper. The hydraulic press is used to apply a uniform pressure of 8 tones/cm^2 on 0.2 g of toner powder to make pellets of 8 mm diameter and 1 ± 0.07 mm thickness. These pellets are also fixed on top of the stub with the help of carbon tape and coated with platinum in a similar manner as that of printed document samples. However, considerable differences in morphological features of pressed pellet toner samples and printed document samples are observed with no appreciable differences in the chemical composition of the toners are seen.

FE-SEM-EDS

The Schottky Field Emission Scanning Electron Microscope Jeol JSM-7610F (Japan) with EDAX detector (AMETEK, USA) is used for imaging and element

detection of samples in the present study. The TEAM™ Software Suite allows spectrum collection and performs characterization of samples. The optimized parameters are as follows:

Magnification	Accelerating voltage	Working distance	Probe current	Low vacuum	Scan rate	Detector
× 1000	15 keV	19 to 20 mm	6 μA	5×10^{-4} Pa	200 s	Secondary electron and backscattered electron, EDAX

The repeatability of this method is studied by analyzing the printed samples from laser printer and photocopier respectively from five different locations. The repeatability of the techniques is determined by exactly superimposed spectra of five scans of a sample as shown in ESM Fig. S1. Similar types of results are obtained for chemical homogeneity testing of printed samples by analyzing one sample five times which resulted in ± 0.002 standard deviation in the weight percent values.

Statistical analysis

Qualitative examination on the dataset is accompanied by multivariate analysis to predict groups of the printed samples having similar spectral properties or elements. However, the information in the datasets is large and complex in nature. Therefore, in order to minimize the difficulty in interpretations, the data in the present research is normalized before analysis. Baseline correction is done by inbuilt SEM software. All statistical analysis is performed by using Microsoft Excel 2010 and IBM SPSS 20.

The use of statistical analysis in the present study involves the extraction of useful information from large datasets by utilizing clustering algorithm, principal component analysis (PCA), and linear discriminant analysis (LDA) statistical methods. These methods construct a model based on the known samples to predict classes for the unknown samples.

Cluster analysis

Cluster analysis (CA) is a descriptive data analysis technique applied to multivariate datasets to uncover the structure present in the data. CA is a valuable tool where no prior knowledge about the dataset is available. It is a type of unsupervised classification, where clusters are formed by evaluating similarities and dissimilarities between the objects occurring in the data. It evaluates the similarity between samples by measuring the distances between them, for instance, samples with similarity will come to lie close to one another forming one cluster, whereas samples with dissimilarity will lie far from each other forming another cluster.

Thus, the cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The similarity between the clusters is calculated by measuring the distance between them, i.e., the shorter distance points towards the large similarity among the clusters. In general, the Euclidean distance is considered as the best choice for the distance metric, because the distances between the samples can be computed directly between two corresponding values. Data clustering can be of either hierarchical or partitioned type. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitioned algorithms determine all clusters at a time.

Hierarchical clustering The hierarchical method (HCA) of clustering follows the reverse procedure where it begins with a single cluster consisting of all observations, forms next clusters, and ends with as many clusters as there are observations. The resultant number of clusters and characteristics of each of them are determined during the analysis. Various distance measures exist to determine which observation is to be appended to which cluster. Different distance metrics can be used to calculate this similarity, the Euclidian distance being the most common and is used in the present study [35–37].

Principal component analysis

Principal component analysis (PCA) is the most extensive unsupervised method for compression and visualization of data. The basic goal of PCA is to retain most of the variation that is present in the given dataset by reducing the dimensions of a large number of interconnected variables to a set of new orthogonal features, called as principal components (PCs). In simple terms, PCA is a method employed for data reduction where new variables are calculated from linear combinations of the original variables. Every new PC in the dataset explains a part of the data variance not described by the previous ones. Thus, the first principal component explains the maximum information from the dataset followed by the second PC and so on. Also, only those PCs whose eigenvalue is > 1 are selected in the final analysis.

This method offers a visual representation of the relationship between samples and variables, as well as an insight into

the way of how the measured values contribute to the similarity respectively to the differences of samples. This makes the PCA technique well suited for multivariate data visualization and interpretation. The original matrix X is decomposed by means of PCA and replaced by T and P. The model has the following equation:

$$X = TP^{\mathrm{T}} \tag{1}$$

In this, T is called scores and has as many rows as the original data matrix and P is called loadings and has as many columns as the original data matrix, and the number of columns in the matrix T equals the number of rows in the matrix P [38–40].

Linear discriminant analysis

The linear discriminant analysis (LDA) is the most widely used supervised pattern recognition technique. The fundamental principle behind LDA is to maximize the ratio of between-class variance and minimize the ratio of within-class variance. LDA helps to build a mathematical function to reduce complex data in the form of variables to new composite dimensions called canonical functions. These canonical functions contain the overall useful information required to predict the separate classes for the samples.

LDA elucidates the dissimilarities among predefined groups of the sample to the greatest extent and develops a model that predicts the group membership of unknown samples [35, 40, 41] based on their characteristics whereas PCA reduces a large number of interconnected variables in a dataset to few new principal components (PCs). Although the new PCs formed are unrelated to each other and explain the variability in the dataset, the maximum separation among the samples is achieved by performing LDA.

Discriminating power

It was first evaluated by Smalldon and Moffat [42] and defined as

$$\mathrm{DP} = \frac{\text{Total number of discriminated sample pairs}}{\text{Total number of possible sample pairs}} \times 100 \tag{2}$$

The idea of using discriminating power is to differentiate the pair of samples on the basis of their qualitative (visual) features. The discrimination in the printouts is achieved by making classes of samples on the basis of differences in their elemental composition as shown in the EDS spectra.

Sample paired t test is also used to check to differences in two closely placed printed samples in the scatter plot. It is a common statistical test used for comparing the means of two independent or paired samples.

Results and discussions

Characterizations of toner from printouts

The FE-SEM-EDS analysis of printed samples reveals the presence of inorganic/organic elements like carbon (C), oxygen (O), aluminum (Al), silicon (Si), calcium (Ca), iron (Fe), zinc (Zn), copper (Cu), sodium (Na), magnesium (Mg), chromium (Cr), potassium (K), manganese (Mn), cobalt (Co), titanium (Ti), nickel (Ni), chlorine (Cl), cesium (Cs), scandium (Sc), and sulfur (S). Some of these elements like C, O, Al, Si, Fe, and Ca are common among most of the samples from various sources but the differences in the weight percent of all these elements are present across samples. The variability in weight percent of the elements is studied to observe the qualitative/quantitative differences in the printouts. The elemental composition in the data is obtained by the system software in the form of weight percent [43] which is the weight of that element measured in the sample divided by the weight of all the elements in the sample multiplied by 100 while utilizing ZAF corrections (where Z = atomic number, A = absorbance, and F = fluorescence). The calculations of the weight percent are done by the system software.

The elements present in the laser and photocopier printed samples are represented in Fig. 1. It shows the distribution of a range of elements in printed documents of both laser printers and photocopier machines. The different colors in this figure (a and b) are used as a measure to depict the presence of elements in accordance with their weight percent in the laser printer and photocopier machines respectively. For example, in Fig. 1a, red color shows the presence of carbon (C) as a dominant element in the samples L18, L20, L22, L26, and L39. Iron (Fe) is the dominant element in the samples L5, L21, L25, L27, and L32 whereas L33 has no carbon and iron as a constituent element but has presence of elements like O (58.78 wt%), Ca (17.84 wt%), Si (11.83 wt%), Al (7.51 wt%), Zn (1.3 wt%), and Cu (2.4 wt%). The laser printer printed samples L24, L31, and L34 have no carbon in their composition but have O as a dominant element with small concentrations of Fe. Similarly, in Fig. 1b, all the samples irrespective of their source of origin contains carbon as a dominant element with an exception (P39) where calcium is acting as a dominant element (55.91 wt%). Although the absence of iron is seen in some samples (P1, P9, P13, P32, P33), two photocopiers printed samples P18 and P19 show the presence of considerable amount of titanium (10.25 and 8.11 wt% respectively). The samples P3, P5, P11, P12, P16, and P18 have carbon, oxygen, and iron as major elements with the presence of other elements like calcium, silicon, magnesium, copper, and zinc in small proportions. The elemental profile of the printouts is dependent on their role in the manufacturing process along with the desired physical and chemical properties of the printed samples.
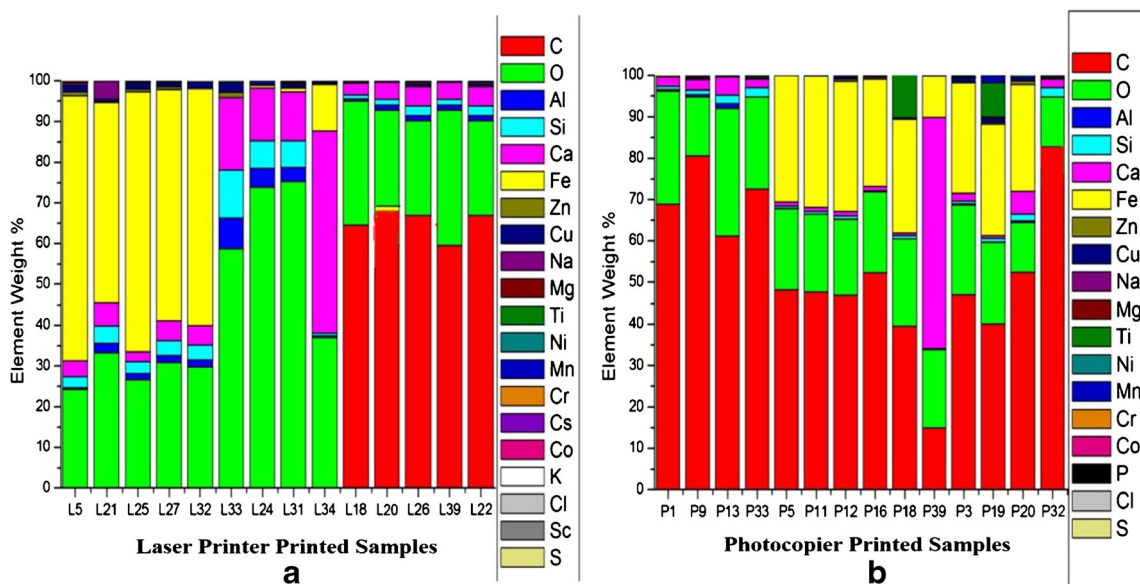
Fig. 1 **a**, **b** Distribution of elemental profile in laser printer and photocopier printed samples respectively

The presence or absence of elements in toners is dependent on the variety of ingredients added in the powder mixture to obtain specific physical and chemical properties. Among the other common constituents, carbon black is added to serve a variety of applications that include coloration, dispersion in resin, and electrical resistivity for charge control properties, suitable particle size distribution, and viscosity for optimum print quality [44]. Besides carbon black, several other materials can be used to make the toners appear black and replace carbon black as a major constituent. Magnetite which has a typical black color is often used to regulate magnetic properties of the toners, but the concentration in which it is added is so high that seldom does the need for any additional pigment arises. Therefore, few toners do not contain any traces of carbon black but have a maximum contribution of magnetite [45].

Few studies have shown the presence of charge control agents like nigrosine as black pigments and their use in toners may replace carbon black effectively. Also, there are other coloring agents like aniline black, furnace black, and thermal black which may be added to impart a black color to the toner. Charge control additives are used for both positive and negative charging applications. For example, quaternary ammonium salts give positive application whereas metal complexes and fumed silica are found to be effective in negative applications [46]. Fumed silica is added to impart multiple properties to the toner such as improved flow properties, hydrophobicity and toner transfer from the photoreceptor to the paper by lowering adhesion, and charge stability between the toner and the carrier mixture.

Other additives like waxes are added to avoid adhesion of the toner to the roll during the process of fusing. The presence of elements like Mg, Zn, and Cu and their combinations thereof may be present in the form of water-soluble metal salts or as an inorganic cationic coagulant to serve as aggregating factors or flocculants. These are added to the toner mixture during preparation at a temperature below the glass transition temperature of the resin or polymer. Blade cleaning is enhanced by blending surfactants and lubricants like zinc stearate, magnesium stearate, and calcium stearate to the surface of the toner [47].

Discrimination of printed documents

Morphological comparison of toners

The samples in the present work are examined with SEM imaging to study the morphological features like size, shape, and distribution of toner particles from different sources. Previously, it has been shown that these features are dependent on the process of manufacturing the toner powders [10]. The most commonly used method is the conventional pulverization but this process creates toner particles of nonuniform sizes and distribution resulting in poor machine performance with a lack of good-quality printouts. With the goal to deliver the desired performance, the toner particles are optimized with more recent methods like chemical polymerization, emulsion aggregation, dispersion polymerization, and chemical milling [48, 49]. The SEM images of some samples from different origins both from the printed sample and dry toner powder pellets are shown in Fig. 2.

This figure depicts the representative SEM images of toner particles both from laser printers and photocopier machines. The contrast in the developed images varies with the differences in the texture of the print. These images are used to study the topographical features of the toner particles like shape, size, and distribution. The shape of the particle in these images varies from regular spheres to irregular spheres with the diameter of the range of ~ 6 to 10 μm as shown in Fig. 2a–
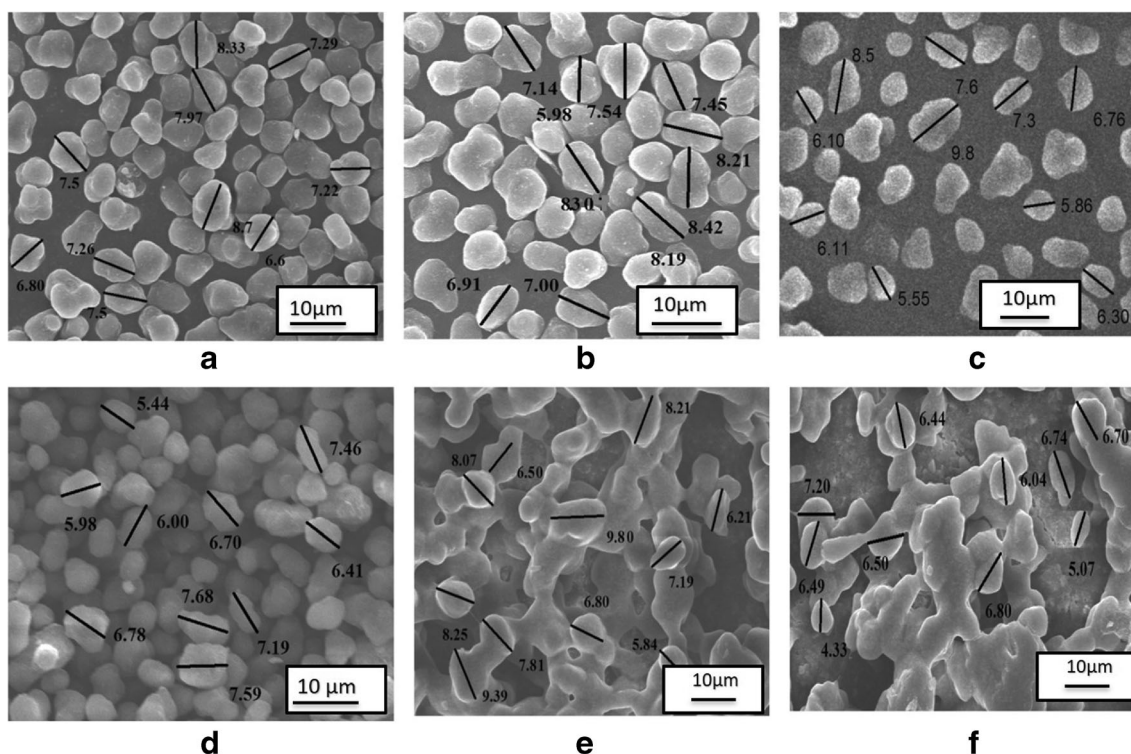
Fig. 2 **a–d** Morphology of toner particles in pressed pellet samples. **e**, **f** Morphology of aggregated particles in the printed document samples

d respectively. It should be noted that the observed average particle size of the photocopier toners is less than that of laser printer toners. The shape and the size play a significant role in determining electrostatic properties, flow performance, and toner adhesion which further affect the quality of printing. The particle aggregation in printed samples is observed when high pressure and temperature are applied to fix the toner powder on the surface of the paper and is shown in Fig. 2e and f. Although all such findings may help the forensic document examiner to differentiate the printed matter either on a single-page or on a few multipage questioned document, in which the morphological differences of the printed toner may point towards the forged nature of the document, however, these findings alone may not allow the examiner to identify the source variability in questioned documents where examination of a large number of printouts is under consideration and hence, such analysis requires some alternate methods.

Statistical classification using paired sample *t* test

In the present study, the morphological differences are observed between laser printer and photocopier printed samples. These morphological features, i.e., shape, size, and the distribution of particles in particular sample, proved useful to differentiate samples between laser printer and photocopier as a source of printing. The mean size of particles in the case of laser printers is found to be more than that of photocopiers and the samples

are further studied with the help of statistical methods to achieve classification among the two different sample sets.

The particles are studied for their mean sizes and classification based on morphological features is achieved utilizing paired *t* test. The sample paired *t* test [50] is used to observe whether any difference is present in the mean values of average particle sizes of laser and photocopier printout toners. Generally, it compares the mean of two related groups of the values to zero. The output of the *t* test is presented in ESM Table S2. The obtained *p* value for both the pairs is 0.00 which resulted in a rejection of the null hypothesis, i.e., both groups have a different particle size of toners after the printing. Again, no correlation ($r = -0.065$) with non-significant *p* value (0.786) is observed in the mean values of these two groups. Thus, laser printer and photocopier printouts are statistically different from each other. Moreover, these groups also show characteristic elemental profiling, and hence, these groups are differentiated morphologically, chemically, and statistically.

However, it is observed that the use of morphological features to classify samples within a same set of samples (within sources of laser printers or photocopiers individually) is challenging due to the presence of particles of different sizes within a same sample.

Elemental profiling of toners from printed documents

Furthermore, the average concentration (weight percent) of each element is calculated and elements like C, O, and Fe

Table 1 Average weight percent of different elements in laser printers and photocopiers in the printed documents included in this study

Laser printer printouts		Photocopier printouts		Inferences
Element	Average wt%	Element	Average wt%	
C	34.22 ± 22.49	C	59.45 ± 15.88	The average weight percent of carbon, oxygen, and iron is more in photocopier printed samples unlike that of laser printer printed samples. The large values of standard deviation in the case of C, O, Ca, Si, and Al elements is because of their absence in some of the printed samples procured from laser printers and photocopier machines.
O	30.61 ± 14.26	O	20.40 ± 5.88	
Al	0.97 ± 1.68	Al	0.38 ± 0.32	
Si	2.30 ± 2.52	Si	1.03 ± 0.55	
Ca	4.89 ± 8.26	Ca	3.92 ± 8.44	
Fe	24.88 ± 17.56	Fe	13.12 ± 12.43	
Zn	0.33 ± 0.39	Zn	0.28 ± 0.24	
Cu	0.73 ± 0.58	Cu	0.48 ± 0.30	
Na	0.17 ± 0.69	Na	0.03 ± 0.07	
Mg	0.08 ± 0.46	Mg	0.07 ± 0.10	
Ti	0.05 ± 0.09	Ti	0.50 ± 2.00	
Mn	0.03 ± 0.07	Ni	0.06 ± 0.19	
Cr	0.09 ± 0.36	Mn	0.17 ± 0.42	
Cs	0.01 ± 0.05	Cr	0.01 ± 0.03	
Co	0.02 ± 0.09	Co	0.01 ± 0.05	
K	0.02 ± 0.7	P	0.01 ± 0.09	
Cl	0.02 ± 0.8	Cl	0.03 ± 0.05	
Sc	0.004 ± 0.02	S	0.01 ± 0.09	
S	0.01 ± 0.07			

are found to contribute maximum towards the toner composition followed by Ca, Si, Al, etc. as shown in Table 1. Although the samples from both the laser printers and photocopiers follow the same trend of average elemental concentration, the amount in each set varies drastically.

The elemental profile of toners is studied in the current research and the obtained spectra of laser printer and photocopier printed documents are shown in ESM Figs. S2 and S3. For example, the common elements in ESM Fig. S2 (a) are C, O, Al, Ca, Si Mg, and Cr; in ESM Fig. S2 (b), samples contain C, O, Al, Ca, Si Mg, and Cr with additional elements like Mn, Na, Ti, and Cu; and ESM Fig. S2 (c) contains one more additional element, i.e., Zn, along with the above-mentioned elements. The similarity in the features of printouts from the same source arises due to the presence of common elements whereas ESM Fig. S2 (d) shows the presence of elements in printed samples from a different source of origin. Similarly, ESM Fig. S3 shows the elements present in photocopier printed samples of the same origin (ESM Fig. S3 (a), (b), and (c)) and from different origins (ESM Fig. S3 (d)). The differences and similarities among samples of the same and different sources can be well studied from these figures. The printed document samples vary in their elemental makeup, hence giving differences to their physical and chemical properties. These differences in the properties are responsible for variations in their surface morphology like particle size, particle shape, and particle size distribution and can be studied by imaging of the printed samples.

Peak to peak comparison

The visual discrimination in the printed samples is analyzed by comparing the presence of elements/peaks at particular energy in all the samples called "peak to peak" comparison. The peak to peak comparison is convenient to perform with the help of a table which indicates all peaks observed in all toner spectra in one frame. The presence/absence of the elements is evaluated and the samples are grouped based on their similarities and differences in the elements as shown in ESM Tables S3 and S4. These groups are only formed in order to predict discrimination among various printed documents on the basis of visual inspection of their spectra obtained by SEM-EDS analysis.

It is shown in ESM Table S3 that among total pairs of laser printer printed samples, 25 sample pairs could not be discriminated because of the large similarity in their visual features (presence/absence of elements). Hence, it is possible that the similar types of chemical composition might be used during the process of manufacturing of the toner powders resulting in similar peak appearances in the printed samples. Similarly, in ESM Table S4, the total set of photocopier toner contains 37 sample pairs that remain non-discriminated by the peak to peak comparison of their spectral features. Therefore, the discriminating power achieved through the Smalldon and Moffatt equation (2) is,

The total number of possible pairs are = 40 (39)/2 = 780 pairs.

As 25 sample pairs have been not discriminated from each other in laser printers, the total number of discriminated sample pairs was = 780–25 = 755 pairs.

The DP for laser printer toners from peak to peak comparison is = 755 × 100/780 = 96.79%.

Similarly, the DP for photocopier toners from peak to peak comparison is = 743 × 100/780 = 95.26%.

Again, the within–same source variability of Hewlett-Packard (HP) laser printer toners and Canon photocopier toners is calculated by Smalldon and Moffatt (2) equations. In Hewlett-Packard laser printer, from 276 pairs of samples, only 13 pairs remain undifferentiated. Similarly, in Canon photocopier toners, from 171 pairs of samples, only 10 pairs remain undifferentiated. Hence, a discriminating power of 95.29% and 94.15% are observed for both HP laser printer toners and Canon photocopier toners respectively by the visual inspection which is highly significant.

Thus, this method delivers overall 96.79% and 95.26% discriminating powers for laser toners and photocopier toners respectively. These results illustrate a good ability of FE-SEM-EDS in the differentiation of printed samples. This method is good for the investigation of smaller sample size. However, for a large group of samples, the peak to peak comparison method has some limitations like time-consuming and tedious and can provide biased results. Moreover, from a forensic perspective, 100% discriminating power is advantageous and could be achieved by applying multivariate analysis. Thus, a method based on statistical modeling is required which provides fast, accurate, and more objective results. In the present research, the HCA algorithm and principal component analysis (PCA) which are more reliable and provide statistical confidence in the outcome are utilized.

Discrimination by multivariate analysis

It is observed that some pairs of printed samples (both from the laser printer toner and from the photocopier toner) are not differentiated by the peak to peak comparison method. Here, the task of discrimination is very challenging and could be possible by utilizing multivariate statistical methods. Therefore, the energy vs. counts/intensity dataset of laser printer and photocopier printed samples from 1 to 15 keV is subjected first to HCA in order to observe any clustering in the print samples. Afterward, the results from HCA are validated by *k*-mean clustering and PCA. This spectral region is selected because it contains the maximum chemical information about the toners. Various preprocessing in the dataset, e.g., normalization and baseline correction, is used to decrease the differences caused by sample preparation and a varying amount of toner during the printing process. The outcomes of HCA analysis are as follows:

HCA algorithm represents an attempt to find good clustering in the dataset using a computationally efficient technique [51]. First, take a closer look at the agglomeration schedule, which displays the clusters combined at each stage and the distances at which this merger takes place. The agglomeration schedule is further used for the estimation of the number of clusters to retain from the data. In the present work, this approximation is estimated by developing a scree plot [51, 52]. In this plot, the significant numbers of the cluster are determined by plotting the number of clusters against the distance at which the printed samples are combined as shown in ESM Fig. S4 (a) and (b). In this plot, a sharp rise (elbow rule) should be noticed in the distance level and it is observed that there is a sharp increase in distance took place at the 29th and 32nd steps of laser printer and photocopier toner respectively. The total number of clusters can be calculated by subtracting the number of stages at which the maximum rise (elbow) in the values is observed from the total number of stages. Thus, the numbers of clusters in the datasets for laser printer and photocopier printed samples are 11 and eight respectively. However, the distance-based decision rule does not work well in every case. The result (number of clusters) is further validated by the dendrogram plot.

In the output of HCA analysis, a tree-like graph, i.e., dendrogram [51], is acquired which basically displays a rescaled distance level at which there is a combination of printouts and hence, clustering took place and is represented in Fig. 3a and b for laser printer and photocopier printed samples respectively. Vertical lines are objects and clusters joined together—their position indicates the distance at which this merger takes place. While creating a dendrogram, SPSS software rescales the distances to a range of 0–25; that is, the last merging step to a one-cluster solution takes place at a (rescaled) distance of 25. The rescaling often lengthens the merging steps, thus making breaks occurring at a greatly increased distance level more obvious. Despite this, it is often difficult to identify where the break actually occurs. Thus, we have started with a large number of clusters and end with one single cluster in the agglomerative approach of HCA.

It is observed that most of the printed samples are grouped into separate clusters and, hence, are distinct from each other. After careful investigation, it is observed that all samples are grouped into 11 and eight different clusters based on the relative agglomerative squared Euclidian distances for laser printer and photocopier printed samples respectively. The results are further validated, e.g., for laser printed documents, by re-analyzing the HCA clustering with a predefined range of cluster solutions from 12 to 10 as observed in the dendrogram. The outcome of this analysis is very promising, especially for 11 segment solutions. In this segment, the laser printed documents with similar spectral appearance are grouped into one cluster whereas the printouts which are relatively unique in its chemical composition are clustered individually. Similar kinds
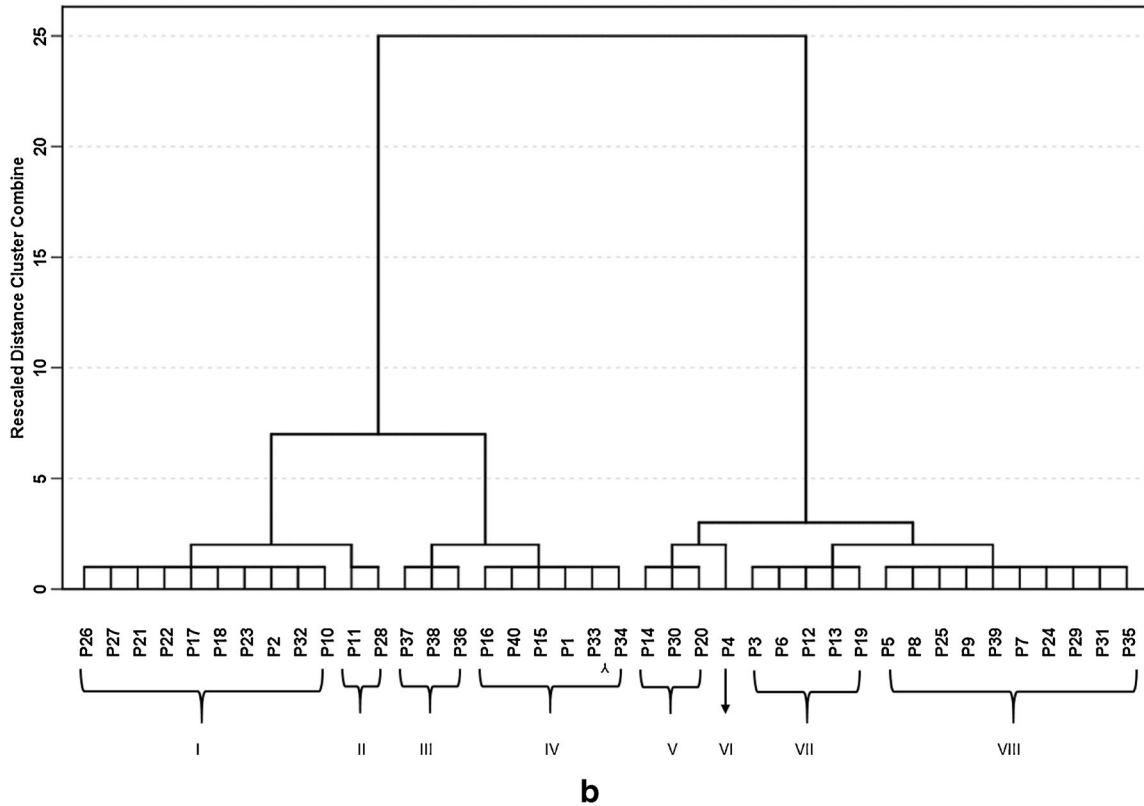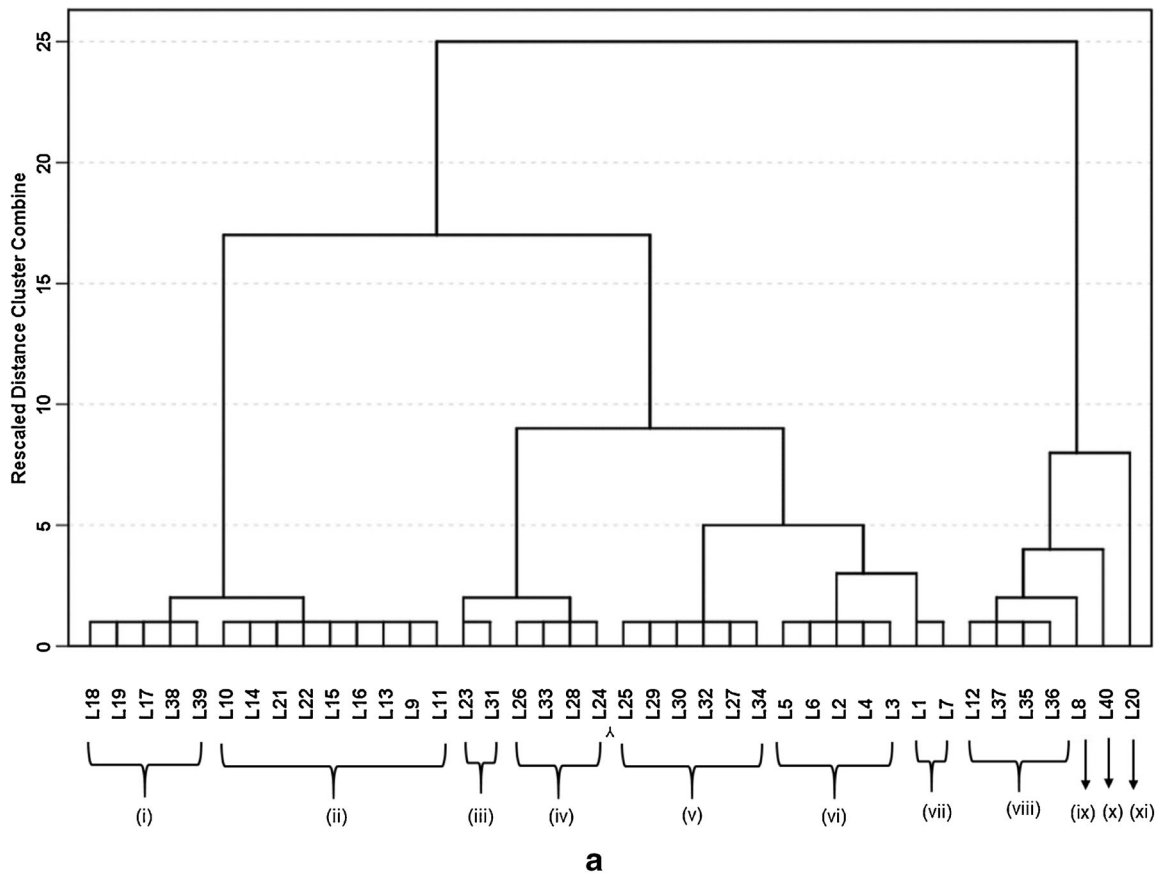
Fig. 3 HCA output for **a** laser printer printed samples and **b** photocopier printed samples

Table 2 Clusters formed by HCA of various printed samples from laser printers and photocopiers. Clusters (i) to (xi) are from laser printers and Clusters I to VIII are from photocopier

Sr. No.	Clusters	Samples	Non-discriminated sample pairs	Discriminating power
1.	(i)	L-17,18, 19, 38, 39	10	89.10%
2.	(ii)	L-9, 10, 11, 13, 14, 15, 16, 21, 22	36	
3.	(iii)	L-23, 31	1	
4.	(iv)	L-24, 26, 28, 33	6	
5.	(v)	L-25, 27, 29, 30,32, 34	15	
6.	(vi)	L-2, 3, 4, 5, 6	10	
7.	(vii)	L-1, 7	1	
8.	(viii)	L-12, 35, 36, 37	6	
9.	(ix)	L-8	0	
10.	(x)	L-40	0	
11.	(xi)	L-20	0	
12.	I	P-2, 10, 17, 18,21, 22, 23, 26,27,32	45	86.92%
13.	II	P-11,28	1	
14.	III	P-36, 37,38	3	
15.	IV	P-1, 15, 16,33,34, 40	15	
16.	V	P-14, 20, 30	3	
17.	VI	P-4	0	
18	VII	P-3,6, 12,13,19	10	
19.	VIII	P-5, 7, 8, 9, 24, 25, 29, 31, 35, 39	45	

of results are obtained for photocopier printed documents. Therefore, combined approaches of all the methodologies confirmed that the datasets of laser printed as well as photocopier printed documents are divided into 11 and eight clusters respectively which might contain similar printed samples based on their chemical compositions.

Further, to ascertain which printed documents grouped under which cluster, k-means clustering is used with predefined groups [51]. This method uses the within-cluster variation as a measure to form homogenous clusters. Specifically, the procedure aims at segmenting the data in such a way that the within-cluster variation is minimized. Based on predefined clusters, the k-means algorithm determines the center for each cluster. Each printed document is then assigned to the cluster center with the shortest distance to it. It is shown in Fig. 3a that in laser printer printed samples, e.g., cluster (i) contained five printed samples, i.e., L17, L18, L19, L38, and L39 and so on. The maximum numbers of samples are grouped into cluster (ii) which contains 9 laser printer samples. All 40 laser printed samples are divided into 11 clusters based on a similar composition of toners in their respective clusters. Similarly, all photocopiers printed samples are divided into eight clusters on the basis of agglomerative distance and might contain similar chemical constituent toners in the individual cluster.

In the case of laser printer printed samples, the total number of non-discriminated sample pairs from all groups is calculated as 85 as summarized in Table 2. It shall be noted here that out of these 85 non-discriminated sample pairs, 32 sample

pairs belong to the common source of origin (HP). For example, cluster (vi) and cluster (vii) contain 5 (L2, L3, L4, L5, L6) and 2 (L1, L7) samples in each respectively and share a common source of origin. The other clusters like (i), (ii), (iv), (v), and (viii) shown in the table have multiple samples but the maximum samples in all these clusters again share their common source of origin. Cluster (iii) contains 2 samples each from a different source. Similarly, the photocopier printed samples have 122 non-discriminated sample pairs distributed among 8 clusters. Cluster II and III have samples belonging to the same source of origin.

It is concluded from the aforementioned discussion that most of the samples get differentiated by using HCA algorithm except some grouped pairs of printed samples in the same cluster. These pairwise samples might contain similar chemical ingredients despite their morphological textures and, hence, remain undifferentiated. Thus, pairwise discrimination for both laser printer and photocopier printed documents provide 89.10% and 86.92% discriminating powers respectively by using a clustering algorithm.

This discrimination is further validated by the PCA method. Prior to the detailed PCA, two tests, i.e., Kaiser-Meyer-Olkin (KMO) and Bartlett tests, are utilized to check the sample adequacy.

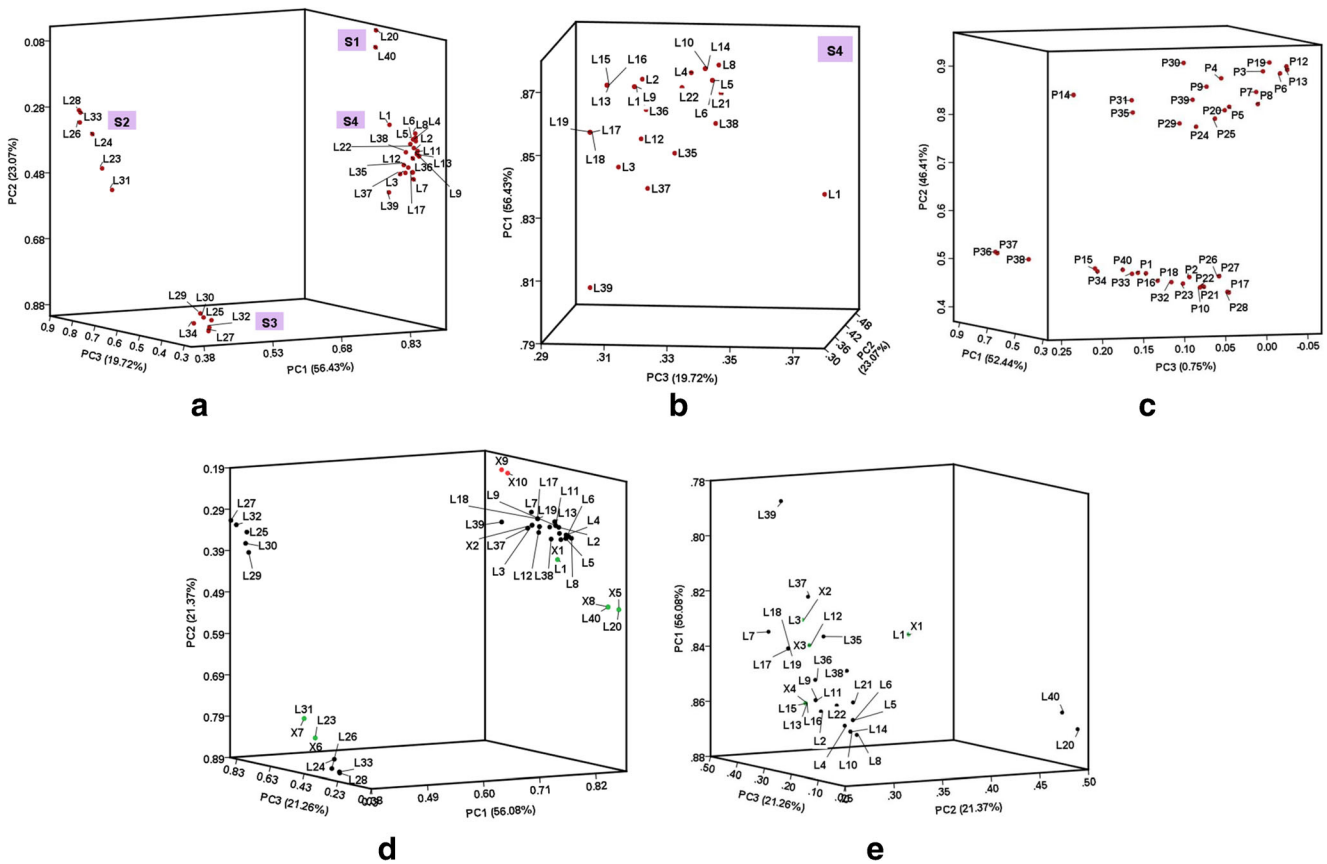(a) Kaiser-Meyer-Olkin (KMO) test: The KMO statistic, which can vary from 0 to 1, indicates the degree to which

Fig. 4 Three-dimensional scatter plot among rotated component values of PC1, PC2, and PC3. **a**, **b** Laser printed document discrimination. **c** Photocopier printed sample discrimination. **d**, **e** Cross-validation of laser printed documents

each variable in a set is predicted without error by the other variables [51]. A value of 0 indicates that the sum of partial correlations is large relative to the sum correlations, indicating factor analysis is likely to be inappropriate. A KMO value close to 1 indicates that the sum of partial correlations is not large relative to the sum of correlations and so factor analysis should yield distinct and reliable factors. This test is performed to show how much the dataset is appropriate for the factor analysis [53, 54].

(b) Bartlett's test: Bartlett's test is an inferential statistic used to assess the equality of variance in different samples [51]. Bartlett's test of homogeneity of variance is based on a chi-square statistic with $(k - 1)$ degrees of freedom, where k is the number of categories (or groups) in the independent variable. This test is suitable to find the correlation between the variables and also used to test the null hypothesis. The variables will be adequate if the p value is less than 0.05. This test is sensitive to check the normality. If the samples came from non-normal distributions, then the Bartlett test provides higher value or non-significant values [53–55].

In the present research, the KMO test shows 0.91 and 0.89 for laser printer and photocopier printed samples and the Bartlett test 0.00 for each respectively which are significant. Hence, this data is adequate for PCA.

Discrimination of laser printer printouts The next step is to check the total variance explained by all the PCs and to determine the significant number of adequate PCs. All the PCs reveal 100% variance in the dataset. However, the first three PCs explain the highest variance, i.e., 99.20% (PC1 = 56.43%, PC2 = 23.07%, and PC3 = 19.72%), present in the dataset. By the fourth PC, the eigenvalue fails to meet the Kaiser Criteria [52]. The remaining 37 components hold only 0.80% of the total variance and hence are not much significant. The similar type of result is obtained by using a scree plot. In this plot, a straight line is obtained after the third PC. As the first three components explain higher variance in the dataset, these PCs are used for discrimination of printed samples via plotting a three dimensional scatter plot among their rotated component values as shown in Fig. 4a.

This figure showed that most of the printouts are differentiated significantly on the basis of their rotated component values into four distinct groups, namely S1, S2, S3, and S4
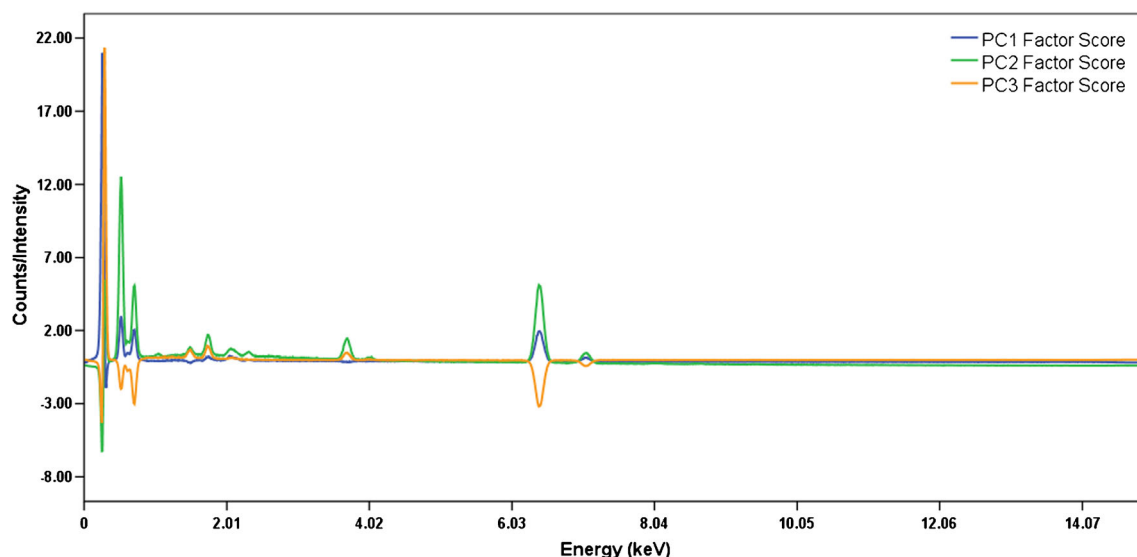
Fig. 5 Regression factor score value plot among PC1, PC2, and PC3 of laser printed documents

respectively. However, some of the printed samples show insignificant differentiation, especially in group S4. When S1, S2, and S3 groups are removed from the final dataset and the resultant data is used to plot a three-dimensional scatter plot of samples belonging to group S4 only, the plot is able to differentiate all samples as shown in Fig. 4b.

Altogether, among all samples, 11 pairs, i.e., L28-L33, L27-L32, L5-L6, L10-L14, L9-L11, L17-L18, L17-L19, L18-L19, L13-L15, L13-L16, and L15-L16, are either superimposed or placed in close vicinity and might share some common chemical constituents and hence show a close resemblance with each other. Among these pairs of samples, nine sample pairs are exactly superimposing with each other (seven pairs, i.e., L5-L6, L9-L11, L10-L14, L17-L19, L13-L15, L13-L16, and L15-L16, having a similar brand but different index number and two sample pairs, i.e., L17-L18, L18-L19, have HP and Samsung origins). These sample pairs might contain similar types of toners or similar chemical constituents in their manufacturing process. All other sample pairs are closely placed and examined by the paired sample t test. It should be noted that some of these samples also show a close similarity in peak to peak comparison method as well as clustering analysis. All other samples are significantly differentiated from each other by multivariate PCA. Therefore, the discriminating power achieved through the Smalldon and Moffatt (2) equation is,

The total number of possible pairs are = 40 (39)/2 = 780 pairs.

As 11 sample pairs have been not discriminated from each other, the total number of discriminated sample pairs was = 780–11 = 769 pairs.

The DP of multivariate analysis comparison is = 769 × 100/780 = 98.59%.

Thus, the discriminating power calculated by multivariate analysis is higher than the discriminating power calculated by the peak to peak comparison and clustering algorithm. The results obtained through PCA are further evaluated by examination of closely placed printed sample pairs via paired sample t test statistics. Basically, t test statistics observed the differences in the mean absorbance values between these pairs. The obtained p value for all three pairs of samples is 0.00 each at 99% confidence interval which resulted in a rejection of the null hypothesis, i.e., both pairs have the same mean absorbance. Thus, the sample L28 is different from L33, and L27 is different from L32. Thus, a combined approach of multivariate analysis and t statistics delivered a 98.85% discriminating power for laser printer printed samples.

Furthermore, to observe the differentiation of printouts with respect to PCs, a rotated component matrix table, i.e., ESM Table S5, has been constructed. It is evident from this table that PC1. PC2, and PC3 are able to differentiate 28, 6, and 6 printed samples respectively as highlighted with bold text. Also, the relationship between PCs and chemical constituents of printed samples is established by developing the regression factor score value plot among PC1, PC2, and PC3 as shown in Fig. 5. This plot reveals that PC1, PC2, and PC3 reflected the peaks arisen at energy 0.28 keV (C), 0.52 keV (O), 0.7 and 6.39 keV (Fe), 1.04 keV (Na), 1.25 keV (Mg), 1.48 keV (Al), 1.73 keV (Si), 2.04 keV (Pt), 2.30 keV (S), 3.69 keV (Ca), 8.04 keV (Cu), and 8.63 keV (Zn). These elements are expected to be closely related to the chemical components of the toner sample, e.g., C is added to make carbon powder or graphite, Fe is used in the form of iron oxide or magnetite, Zn and Mg form zinc stearate and magnesium stearate, and Si forms fumed silica and silicon oil along with other elements that act as metal complexes, ammonium salts, nigrosine, azo pigments, polyethylene wax, aniline black, etc.

However, the exact chemical composition is a kept trade secret and, hence, cannot be predicted. All these elements contribute majorly to the toners by providing coloring, electrostatic properties, hydrophobicity, flowability, and lubrication.

At energy 0.28 keV, PC2 opposes PC1 and PC3 which depict the presence of carbon black. Similarly, PC3 opposes PC1 and PC2 at energies 0.52 keV, 0.7 keV, and 6.39 keV which correspond to the oxides of iron, sulfur (sulfoxides), or carbon (carbonates). PC1 opposes PC2 and PC3 at 1.48 keV and 3.69 keV which depict the peak of aluminum and calcium respectively. Moreover, only PC2 shows the peaks of copper and zinc elements at their respective energies. Some minor peaks of platinum are also observed and favored by all three PCs. This discussion explains the distributions of chemical constituents in laser printers with respect to important principal components by analyzing their extracted regression factor score values.

Discrimination of photocopier printouts In the photocopier printed samples, the first three PCs explain the highest variance, i.e., 99.61% (PC1 = 52.44%, PC2 = 46.42%, and PC3 = 0.75%), present in the dataset. However, by the fourth PC, the eigenvalue fails to meet the Kaiser criteria. The similar type of result is obtained by using a scree plot. As the first three components explain higher variance in the dataset, these PCs are used for discrimination of printed samples via plotting a three dimensional scatter plot among their rotated component values as shown in Fig. 4c.

This figure showed that most of the samples are differentiated significantly on the basis of their rotated component values. However, some sample pairs that superimpose completely (same source of origin with different index number) or lie very close (different origins) to each other are expected to share common chemical constituents. The sample pairs P12-P13, P10-P21, P10-P22, P21-P22, P17-P28, P26-P27, P15-P34, and P36-P37 might share some common chemical constituents and might not be differentiated. Among these pairs of samples, four sample pairs are exactly superimposing with each other having a similar brand but different index number and one sample pair, i.e., P26-P27, has Canon and Konica Minolta origins. Furthermore, all the closely placed sample pairs are examined by the paired sample t test. Other photocopier printed samples are differentiated from each other by applying multivariate PCA. Therefore, discriminating power calculated is as follows:

> Eight sample pairs have been not discriminated from each other. Thus, the total number of discriminated sample pairs is = 780–7 = 772 pairs.
> The DP of multivariate analysis comparison is = 772 × 100/780 = 98.97%.

Thus, the discriminating power calculated by multivariate analysis is higher than the discriminating power calculated by the peak to peak comparison and clustering methods.

The results obtained through PCA analysis is further evaluated by examination of closely placed sample pairs, i.e., P12-P13, P10-P21, P10-P22, and P15-P34, via paired sample t test statistics. The obtained p value for these pairs of the samples is 0.00 at a 99% confidence interval which resulted in a rejection of the null hypothesis, i.e., all sample pairs have the same mean counts/intensity. Thus, the sample P12 is different from P13, P10 is different from P21, P15 is different from P34, and so on. Thus, a combined approach of multivariate analysis and t statistics delivered a 99.49% discriminating power for photocopier printed samples.

The differentiation of photocopier printed samples with respect to PCs is evaluated by analyzing the rotated
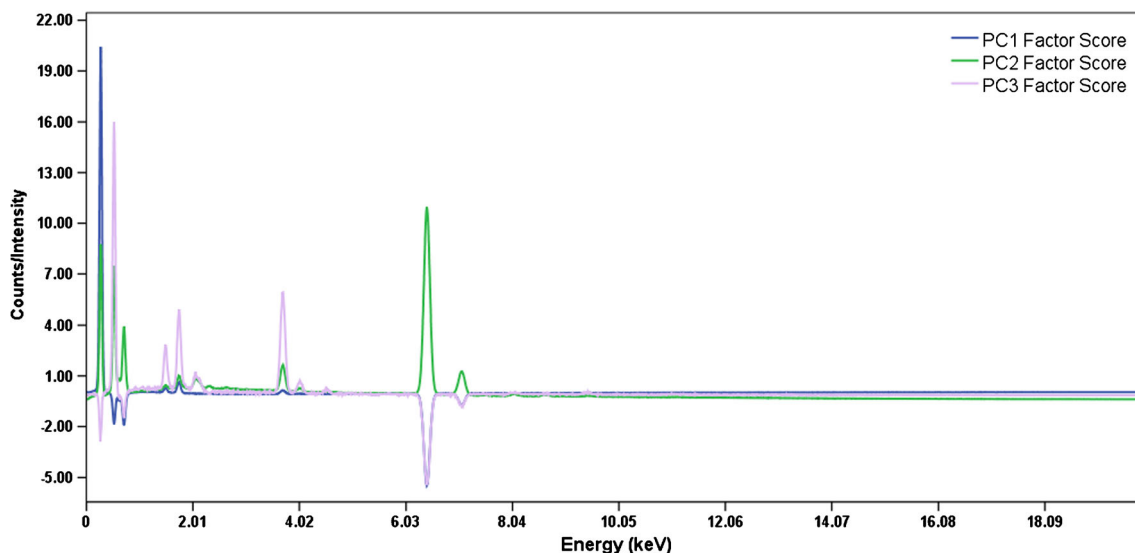

Fig. 6 Regression factor score value plot among PC1, PC2, and PC3 of photocopier printed documents

component matrix values. It is observed that PC1, PC2, and PC3 are able to differentiate 21, 18, and one printed samples respectively. Furthermore, the relationship between PCs and chemical constituents of printed samples is established by developing the regression factor score value plot among PC1, PC2, and PC3 as shown in Fig. 6. This plot also focuses on the major elements present in the photocopier printed samples. It is evident from the figure that at energy 0.28 keV, PC3 opposes PC1 and PC2 which depict the presence of carbon black. Similarly, PC1 opposes PC2 and PC3 at energy 0.52 keV which corresponds to the presence of different elemental oxides. PC1 and PC3 oppose PC2 at 0.7 keV and 6.39 keV which depict the peak of oxides of iron, sulfur, carbon.

Cross-validation of the PCA model The practical application of this approach is also studied by analyzing ten unknown randomly selected laser printer printed samples, i.e., X1, X2, X3, X4, X5, X6, X7, X8, X9, and X10. The prediction accuracy, i.e., to which groups these samples belong, is tested by employing the same methodology as used in PCA analysis. After collecting the SEM-EDS spectra of all ten unknown samples, the normalized datasets are subjected to PCA model along with all known laser printer printed samples. The first three PCs are enough to describe 98.71% of total variance present in the dataset. Furthermore, a three-dimensional scatter plot is developed among the rotated component values of PC1, PC2, and PC3 as represented in Fig. 4d.

It is seen that after the removal of differentiated samples from the original dataset and plotting a three-dimensional scatter plot of only grouped samples, all the samples are classified as shown in Fig. 4e. It is evident from both the scatter plots that eight samples from all unknown samples got superimposed with their respective printed samples, i.e., X1 belongs to printer L1, X2 belongs to printer L3, X3 belongs to printer L12, and so on. However, two unknown printed samples, i.e., X9 and X10, do not belong to any of the mentioned laser printers (Table 1) and are placed separately. It signifies that eight unknown samples are procured from the same printer as mentioned in Table 1 and two samples, i.e., X9 and X10, are procured from some other source of printing. After verification of the unknown samples, it is found that the actual origin of these samples is the same as obtained in our study. Therefore, the presented approach predicted the source of origin of unknown printed samples.

Classification of printed documents

The classification of the printed samples is necessary for analysis of unknown printed documents. It will help in reducing the chance of false positive results effectively. After organizing the samples in clusters, further analysis is done by LDA to classify the printed samples on the basis of elemental spectra.

In the predictive analysis, the variables selection plays a crucial role. The counts/intensity values of printed are utilized samples as they are directly related to the weight percent of elements and, hence, the concentrations. The concentrations of these elements vary from toner to toner in each source of the laser printer and photocopier and hence are used in the classification model. All the printed samples give peaks of different elements at specific energies which depict the elemental composition of toners. Thus, the counts/intensity values of C, O, Al, Si, Ca, Fe, Zn, Cu, Na, Mg, Ti, Ni, Mn, Cr, Co, K, Cl, Sc, and S at particular energies are selected for LDA modeling and are used as the independent variable and "groups" is used as the dependent variable. These energies represent the maximum intensity of various elemental ingredients in toners.

Analysis of variance test

This test allows us to measures each independent variables potential before the model is created. Each test displays the results of a one-way analysis of variance (ANOVA) for the independent variable using the grouping variable as the factor. If the significance value is greater than 0.10, the variable probably does not contribute to the model.

Wilks' lambda is used to test whether there are differences between the means of recognized sample groups based on the dependent variable combination. It is a direct measure of the proportion of the variance that is unaccounted for by the independent variable. This test allows for selecting the best predictor of the grouping variable. For the present study, the value of Wilks' lambda is 0.81 with p value 0.000, i.e., < 0.05. Also, Box's M is used to test the homogeneity of covariance matrices based on the likelihood test ratio and is also showing the significant p value = 0.00. Among all the variables entered in classification software, only the counts/intensity values of Fe, Al, O, and C have been entered in the final model. These elements are selected by the discriminant analysis model itself because the selected elements are more significant than other elements.

Canonical discriminant function coefficient

The unstandardized coefficients are used to create the discriminant function (DF) equation. A prerequisite condition of eigenvalue > 1 and canonical correlation > 0.35 should be followed to develop a good model. The test gives eigenvalue of 11.28 which is greater than 1 and the canonical correlation of 0.95 which is also greater than 0.35. Thus, the obtained equation explains the grouping of printed samples well. The equation for grouping the printed samples is taken in consideration w.r.t major elements and is calculated as:

$$DF\ Equation = -21.04 + 10.505[Fe] + 103.814[Al]$$
$$+ (-4.301)[O] + 1.350[C] \qquad (3)$$

Table 3 Discriminant function value of blind samples studied by cross-validation

Sample	Carbon (E = 0.28 keV)	Oxygen (E = 0.52 keV)	Aluminum (E = 1.49 keV)	Iron (E = 6.40 keV)	Score discriminant value	Inference
B1	3.371	1.417	0.105	0.574	−5.656	Photocopier machine
B2	2.218	1.025	0.118	1.205	2.449	Laser printer
B3	2.996	0.894	0.100	1.035	0.452	Laser printer
B4	2.387	0.914	0.109	1.197	2.135	Laser printer
B5	2.271	1.534	0.181	0.876	3.415	Laser printer
B6	2.651	1.356	0.175	0.598	1.156	Laser printer
B7	3.262	1.677	0.156	0.740	−0.585	Photocopier machine
B8	2.651	1.356	0.313	0.598	15.429	Laser printer
B9	7.191	0.554	0.209	0.107	9.10	Laser printer
B10	6.538	1.034	0.181	0.030	2.444	Laser printer

Equation (3) clearly indicates that Al plays a significant role in the discrimination of laser printed samples because it shows maximum coefficient values among all elements. Similarly, O also plays an important role in the discrimination for its negative coefficient values. A further way of interpretation of discriminant analysis results is to describe each group in terms of its profile, using the group means of predictor variables. The group means are called centroid. Cases with scores near to a centroid are predicted belonging to that group.

For practical purposes, a cut score is calculated which is halfway between the two centroids:

$$\text{Cut score} = (-3.187 + 3.187)/2 = 0$$

Therefore, if the score of the discriminant function is greater than 0, the printed document will belong to the laser printer group, and if the discriminant function value reduces from 0, it is said to belong to the photocopier group. Furthermore, the results are validated with leave-one-out cross-validation methodology.

Classification results

The classification studies are performed by including the 20 representative samples for validation purposes. The samples from a different source of origin (all brands) are incorporated along with some replicates (HP, Brother, Samsung, Canon, Xerox, and Konica Minolta) from the laser printer and photocopiers printed samples respectively. The software validation has been done on all 20 samples.

On the basis of all observations and model "goodness of fit," it is concluded that the use of counts/intensity values as a variable for LDA resulted in the original classification of 100% for both the datasets. Thus, the original classification

for both the laser printer and photocopier shows good predictive model. Furthermore, the leave-one-out cross-validation result shows a combined 100% classification of printed samples which is very significant. It is expected that the developed model would predict the membership of unknown questioned samples accurately.

Cross-validation of LDA model

The validation of the developed classification model in the present study is performed by analyzing the printed samples of unknown origin to the author. For this, ten blind samples from different laser printers and photocopiers are collected and analysis is performed on developed discriminant function equations. These samples are named as B1, B2, and so on. The normalized counts/intensity values of each element at their specific energies are put in Eq. (3) to obtain the score of discriminant function as shown in Table 3. After the cross-validation, it is found that a negative score discriminant function shows that the samples belong to photocopier, whereas the positive value concludes that the sample is obtained from a laser printer. All samples are correctly classified in their respective class.

Thus, two statistical models, i.e., PCA and LDA, are developed along with FE-SEM-EDS database of all collected printed samples through which it is convenient to identify the source of origin of unknown printed samples. After verification of these unknown samples, it is found that the actual origin of the sample is the same as that obtained in our study. Therefore, 100% of classification results are achieved. Here, it should be noted that the significant classification results using the current developed models are best achieved only after the normalization of obtained FE-SEM dataset. This is done to avoid any biased results.

Conclusions

In this study, FE-SEM-EDS analysis is performed on printed documents obtained from laser printers and photocopiers in a quasi-nondestructive manner with a high resolution, accuracy, reliability, and repeatability. The SEM images of the printed document as well as toner pellets reveal the characteristic differences between the morphology of toner particles and their distribution. All these features are dependent on the mode of preparation which is again kept private with the manufacturer. The printouts are studied for the presence and absence of elements and their contribution to the nature of the toners. Elements like carbon (C), oxygen (O), aluminum (Al), silicon (Si), iron (Fe), magnesium (Mg), sodium (Na), titanium (Ti), calcium (Ca), copper (Cu), and zinc (Zn) are commonly found in the EDS spectra of both laser printers and photocopier machines with variations in the weight percent of each element. Among all these samples, C, O, Al, and Fe are found to be the dominant elements followed by Ca, Si, Zn, etc. All these elements are added to impart particular properties that remain identifiable to their origin.

After characterization, the discrimination of printed samples is achieved by four different methods, i.e., morphology comparison, peak to peak comparison, clustering algorithm, and PCA and t test analysis. The peak to peak comparison method delivers 96.79% and 95.26% of discriminating power for laser printer and photocopier printed samples whereas clustering algorithm provides 89.10% and 86.92% of discriminating power for laser and photocopier printed samples respectively. The maximum discrimination is provided by a combined approach of multivariate PCA and t test statistics which resulted in 98.85% and 99.49% discrimination powers for laser printer and photocopier printed samples respectively. More importantly, a discriminant model is developed to classify the unknown printed sample to its respective groups. The model based on linear discriminant analysis reveals 100% accurate grouping of the printouts by leave-one-out cross-validation approach.

The practical application of this approach is studied by analyzing the source of origin of ten unknown laser printer printed samples by PCA and LDA models. It is clearly indicated from the scatter plot in the PCA model that eight unknown samples get superimposed with their respective printer sample whereas two samples do not belong to any of the laser printer investigated in the present study. Similarly, from all ten blind samples in LDA models, two samples, i.e., samples B1 and B7, belong to the photocopier machines whereas, the rest of the eight samples belong to the printouts from laser printers.

Therefore, the presented approach predicted the unknown printed document group accurately. However, the fundamentals knowledge and statistical assumptions should be followed before applying such methods; otherwise, biased results might be obtained. Again, the scope of the questioned document is wide open particularly in the existence of particular ink/toners in particular year by employing collective modern analytical and chemometric methods. Generally, the chemical makeup or elemental profile of the toner powders is modified year by year to enhance the quality of printing by adding or replacing one or more components with new or different chemical formulations. These changes being time-dependent will allow the examiner to ascertain the authenticity of any questioned document by studying its features based on the availability or existence of the particular toner in that particular period. Thus, the present approach can be utilized in such cases.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Kelly JS, Lindblom BS. Scientific examination of questioned documents. 2nd ed. New York: Taylor Francis; 2006.
2. Marshall G. Recent progress in toner technology. Soc Imaging Sci Technol; 1997.
3. Scharfe M. Electrophotography and optimization. Letchworth, Hertfordshire, England: Research Studies Press Ltd. New York: John Wiley & Sons INC; 1984.
4. James EL. The classification of office copy machines from physical characteristics. J Forensic Sci. 1987;32:1293–304.
5. Totty R. The examination of photocopy documents. Forensic Sci Int. 1999;46:121–6.
6. Kipphan H. Handbook of print media: technologies and production method: Springer; 2001.
7. Schien LB. Recent advances in our understanding of toner charging. J Electrost. 1999;46:29–36.
8. Leach R. The printing ink manual. 4th ed. Berlin: Springer Verlag Gmbh; 1988.
9. Park J, Ahn S, Yoo J, Kim D, Kim DS. Effects of polymerization process variables on the properties of suspension polymerized toner. In: 18th International Conference on Composite Materials, Edinburgh, Scotland; 2009.
10. Yang J, Wang TJ, He H, Wei F, Jin Y. Particle size distribution and morphology of in situ suspension polymerized toner. Ind Eng Chem Res. 2003;42:5568–75.
11. Lennard C, Mazzella W. A simple combined technique for the analysis of toners and adhesives. J Forensic Sci Soc. 1991;31(3):365–71.
12. Meng P, Wang S. Identification of copy machines through pyrolysis chromatography of toners. Int J Doc Exam. 1995;5:88–90.
13. Verma N, Kumar R, Sharma V. Analysis of laser printer and photocopier toners by spectral properties and chemometrics. Spectrochim Acta A Mol Biomol Spectrosc. 2018;196:40–8.

14. Merrill RA, Bartick EG, Taylor JH. Forensic discrimination of photocopy and printer toners, part II. Anal Bioanal Chem. 2003;376:1279–85.
15. Materazzi S, Risoluti R, Pincia S, Romolob FS. New insights in forensic chemistry: NIR/chemometrics analysis of toners for questioned documents examination. Talanta. 2017;174:673–8.
16. Udristioiu EG, Bunaciu AA, Aboul-Enein HY, Tanase I. Infrared spectrometry in discriminant analysis of laser printer and photocopy toner on questioned documents. Instrum Sci Technol. 2009;37:230–40.
17. Trzcinska BM, Brozek-Mucha Z. The possibilities of identifying photocopy toners by means of infrared spectroscopy (FTIR) and scanning microscopy. Mikrochim Acta. 1997;14:235–7.
18. Trzcinska BM. Classification of black powder toners on the basis of integrated analytical information provided by Fourier transform infrared spectrometry and X-ray fluorescence spectrometry. J Forensic Sci. 2006;51:919–24.
19. Egan WJ, Morgan SL, Bartick EG, Merrill RA, Taylor JH III. Forensic discrimination of photocopy and printer toners. II. Discriminant analysis applied to infrared reflection–absorption spectroscopy. Anal Bioanal Chem. 2003;376:1279–85.
20. Trzcinska BM. Discriminating among powder toners of the same colour by the use of spectroscopic methods. Probl Forensic Sci 2015;80–95.
21. Johnson C, Martin P, Roberts K, Trejos T, Corzo R, Almirall J, et al. The capability of Raman micro spectroscopy to differentiate printing inks. J Forensic Sci. 2017:1–14.
22. Chu PC, Cai BY, Tsoi YK, Yuen R, Leung SK, Cheung NH. Forensic analysis of laser printed ink by X-ray fluorescence and laser-excited plume fluorescence. Anal Chem. 2013;85:4311–5.
23. Williamson R, Raeva A, Almirall J. Characterization of printing inks using DART-Q-TOF-MS and attenuated total reflectance (ATR) FTIR. J Forensic Sci. 2016;61:706–14.
24. Trejos T, Torrione P, Corzo R, Raeva A, Subedi K, Williamson R, et al. A novel forensic tool for the characterization and comparison of printing ink evidence: development and evaluation of a searchable database using data fusion of spectrochemical methods. J Forensic Sci. 2016;61:716–24.
25. Egan WJ, Galipo RC, Kochanowski BK, Morgan SL, Bartick EG, Miller ML, et al. Forensic discrimination of photocopy and printer toners. III. Multivariate statistics applied to scanning electron microscopy and pyrolysis gas chromatography-mass spectrometry. Anal Bioanal Chem. 2003;376:1286–97.
26. Gilmour CL. A comparison of laser printed and photocopied documents can they be distinguished? J Can Soc Forensic Sci. 1997;27:245–59.
27. Brandi J, James B, Gutowski SJ. Differentiation and classification of photocopier toners. Int J Forensic Doc Exam. 1997;3:324–44.
28. Trejos T, Corzo R, Subedi K, Almirall J. Characterization of toners and inkjets by laser ablation spectrochemical methods and scanning electron microscopy-energy dispersive X-ray spectroscopy. Spectrochimica Acta B. 2014;92:9–22.
29. Corzo R, Subedi K, Trejos T, Almirall J. Evaluation of the forensic utility of scanning electron microscopy-energy dispersive spectroscopy and laser ablation-inductively coupled plasma-mass spectrometry for printing ink examinations. J Forensic Sci. 2016;61:725–34.
30. Szynkowska MI, Czerski K, Paryjczaka T, Parczewskib A. Ablative analysis of black and colored toners using LA-ICP-TOF-MS for the forensic discrimination of photocopy and printer toners. Surf Interface Anal. 2010;42:429–37.
31. Almirall J, Trejos T. Applications of LAICPMS to forensic science. Elements. 2016;12:335–40.
32. Subedi K, Trejos T, Almirall J. Forensic analysis of printing inks using tandem LIBS and LAICP-MS. Spectrochim Acta B At Spectrosc. 2015;103:76–83.
33. Lennard C, Deftar M, Robertson J. Forensic application of laser-induced breakdown spectroscopy for the discrimination of questioned documents. Forensic Sci Int. 2015;254:68–79.
34. Metzinger A, Rajko R, Galbacs G. Discrimination of paper and print types based on their laser induced breakdown spectra. Spectrochim Acta B. 2014;94–95:48–57.
35. Kumar R, Sharma V. Chemometrics in forensic science. TrAC Trends Anal Chem. 2018;105:191–201.
36. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufmann L. Chemometrics textbook. In: Elsevier; Amsterdam; 1988.
37. Møller SF, von Frese J, Bro R. Robust methods for multivariate data analysis. J Chemom. 2005;19:549–63.
38. Brereton RG. Chemometrics: data analysis for the laboratory and chemical plant. Chichester: Wiley; 2003.
39. McLachlan G. Discriminant analysis and statistChichesterical pattern recognition: Wiley; 2004.
40. Sharma V, Kumar R. Trends of chemometrics in bloodstain investigations. TrAC Trends Anal Chem. 2018;107:181–95.
41. Huberty CJ, Olejnik S. Applied MANOVA and discriminant analysis. 2nd ed. Chichester: Wiley; 2007.
42. Smalldon KW, Moffat AC. The calculation of discrimination power for a series of correlated attributes. J Forensic Sci Soc. 1973;13:291–5.
43. Kursula PK. Accuracy, precision and detection limits of SEM-WDS, SEM-EDS and PIXE in the multi elemental analysis of medieval glass. X-Ray Spectrom. 2000;29:111–8.
44. Gruber RJ, Julien PC, Toner Technology D. Handbook of imaging materials. 2nd ed. New York: Marcel Dekker; 1991.
45. Ataeefard M, Ghasemi E, Ebadi M. Effect of micro and nano magnetite on printing toner properties. Sci World J. 2014;7:706367.
46. Birkett KL, Gregory P. Metal complex dyes as charge control agents. Dyes Pigments. 1986;7:341–50.
47. Hartus T. Adhesion of electrophotographic toner on paper. Graphic Arts Finland 2001;30:3.
48. Nienow AW, Ding P, Pacek AW, Abinhava K, Pickard S, Edwards MR. A process for the manufacture of chemically produced toner (CPT). II. Effect of operating conditions. Ind Eng Chem Res. 2005;44:6012–21.
49. Ataeefard M. Production of carbon black acrylic composite as an electrophotographic toner using emulsion aggregation method: investigation the effect of agitation rate. Compos Part B. 2014;64:78–83.
50. Kim TK. T test as a parametric statistic. Korean J Anesthesiol. 2015;68:540–6.
51. Sharma V, Bharti A, Kumar R. On the spectroscopic investigation of lipstick stains: forensic trace evidence. Spectrochim Acta A Mol Biomol Spectrosc. 2019;215:48–57.
52. Anandarajan M, Hill C, Nolan T. Cluster analysis: modeling groups in text. In: Practical text analytics, advances in analytics and data science, vol. 2. Cham: Springer; 2019.
53. Kaiser HF. An index of factorial simplicity. Psychometrika. 1974;39:31–6.
54. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate data analysis. Upper Saddle River: Pearson Prentice Hall; 2006.
55. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7th ed. Upper Saddle River: Prentice Hall; 2010.