The classification of pen ink aging by machine learning and deep learning technique using Raman spectrum

Received: 26 January 2023; Accepted: 8 March 2023 *Research Article*

Kübra GÜRBÜZ GÖÇMEN
Engineering Faculty
Istanbul Ticaret University
Istanbul/Turkey
kubra.gocmen@istanbulticaret.edu.tr

Mustafa Cem KASAPBAŞI
Engineering Faculty
Istanbul Ticaret University
Istanbul/Turkey
mckasapbasi@ticaret.edu.tr

Sinan BOSNA
Physics Engineer
Tübitak Bilgem
Kocaeli/Turkey
sinan.bosna@tubitak.gov.tr

Abstract—Forgery of valuable documents generally constitutes falsification methods based on altering a previously written document by using similar or identical ink. In the event of the aforementioned situation, forensic science experts conduct various technical examinations on the relevant document using different devices. One of the main purposes of these examinations is to determine the differences in the aging levels of the inks relative to each other. Raman Spectra, which is also used for different purposes in forensic sciences, is one of the methods that can be used in this field. The Raman spectrometer provides information about molecules' vibration energy levels and presents the analyzed region's spectral signature values. Experts can observe the time-dependent changes that occur in the substances in the region under investigation relative to each other and in the substance content through the information obtained. Utilizing this information, sample data were created at different times using the same pen on A4 paper in our study. These data were divided into two groups old and new data. Raman spectra were taken with a 785 nm laser on both sample data. Sequential Keras model, KNN, and SVM algorithms were used to detect ink aging on paper. The k-fold cross-validation method was used to determine the classification performance more accurately. The results showed that the classification performance was 98.71% for the neural network and 100% for the KNN and SVM.

Keywords—Raman spectroscopy, pen ink, machine learning ,deep learning

I. INTRODUCTION

Any modification to a forensic document is regarded as document forgery because it does not accurately reflect the facts. There are falsification methods such as changes, additions and scribbling on the document. When a suspicious document is encountered, it is handled as a forensic case and examined by a document examiner. A document examiner is a person who specializes in studying and researching to uncover the facts about documents. The document "examiner" should not only specialize in handwriting, typewriting and printer printouts, but also in forgery, paper and ink analysis, falsified documents and all technical devices and methods used in document preparation [1]. Forged documents differ from the originals in a number of ways (print quality, paper structure, dimensions, typeface, characteristic features in numbers and patterns) [2]. Document review devices have been developed to detect these differences.

Raman scattering is a type of inelastic scattering. The spectrometric analysis of these scatterings, which occur as a result of the change in the vibration energy modes of molecules as a result of the interaction of matter and light, provides information about the bonds and structures of molecules. In order to obtain the Raman spectrum, a laser source, a spectrophotometer and optical elements that will optically block the laser source in the spectrophotometer and focus the laser source on the sample.

Raman spectroscopy is used in many fields to identify unknown substances, verify samples in quality assurance, analyze the chemical composition of samples or monitor changes. In addition, in recent years, Raman spectra have been used in applications such as classification and separation of certain analytes, disease detection from blood serum, etc [3] [4]. In this study, it is aimed to discriminate whether the pen ink is freshly written or not by taking Raman spectra over the ink sample. For this purpose, data were collected from different people using the same pen on white A4 paper with an interval of approximately 5 years. The spectrum of the ink on the data obtained was taken using Raman spectroscopy. The peaks of the spectra were equalized using min-max normalization and baseline correction. Raman spectra have a strong background fluorescence so that baseline correction is important. There are many methods for minimizing the fluorescence background signal. Various methods were evaluated and morphology based baseline correction method was used [5]. Machine learning algorithms KNN and SVM were used. Sequential model was used using Keras library, one of the deep learning algorithms. The results obtained from machine learning algorithms were more successful than neural network. [6]

II. METARIALS AND METHODS

A. Preparation of Working Area

Raman spectra were performed using Ocean Insight Raman QE Pro High spectrometer, 785 nanometer (40 mW) laser and optical probe. Samples were placed at the focal length of the probe (10 mm) and recorded in the dark environment with 500 ms integration time and 5 averages. The papers were fixed to the floor with the help of weights during the measurement to avoid focal length changes.

B. Data acquisition

About five years ago, data were collected from different people using a Schneider Xtra 8053 pilot pen on A4 paper. The data collection sheet consists of nine sections. The first section contains the name, surname, age, hand direction and date of the person from whom we collected data. In the second section, numbers between 0-9 were printed. In the third section, the phone number, in the fourth section, the name and surname information was written again, in the sixth section,

the alphabet was written with upper and lower case letters, and in the other sections, Fig. 1 in order of priority. These data were collected extensively to be used for future studies. For this study, Raman spectroscopy measurements were taken in the dark using the upper right corner of the drawn arrow shape.

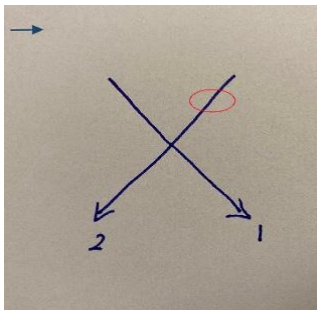

Fig. 1: Measurement sample data

The current data were collected using the same format and including similar or different people who wrote the experiment papers taken about five years ago. The pen used five years ago was the same pen used for the new data collection. The experimental pen was kept throughout the process. The data collected five years ago were kept in files in a cabinet in the office environment. When Raman spectroscopy measurements were taken with the data left on the desk for five years in the office environment, spectral differences were observed compared to the measurements taken from the papers kept in the cabinet for five years. It was observed that ink aging was visibly different on the papers kept outside. Sample data was generated on A4 paper with the current Schneider Xltra 8053 pilot pen. Raman spectra were taken of the data generated from the pilot pen used five years ago and the newly purchased pilot pen. When the spectra were plotted, it was observed that the spectra of the old and new pen were similar.

C. Classification

A total of 155 data, old and new, were collected from different people. Raman measurements were taken at the same location on each paper and spectrum data were obtained. There were negative values in the spectrum data. For each data, spectrum values corresponding to the same wavelength were deleted. Min-max normalization was performed to bring the peaks of the spectral values to the same point. Min-Max normalization is used to scale values in the range (0,1). It sums the data between (0,1) while preserving the values of the original data.

$$x_s = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

III. THEORY

Supervised machine learning methods are often used for classification and regression problems. Supervised machine learning involves labeled data that gives information about what is in the dataset. Since it is known what the data corresponds to, the algorithm can give the appropriate output when a similar test data is presented. In unsupervised machine learning, it makes inferences based on the similarity of the datasets to each other.

Deep learning is a type of machine learning that uses an artificial neural network model to learn on large and complex data. It can analyze complexity in text, audio and image data. It is frequently used in problems such as translating audio files into text, extracting meaning from images, and detecting writing characteristics.

A. Support Vector Machines (SVM's)

SVM is a supervised learning method and is often used for classification and regression problems. It separates two classes by a plane. The plane is determined according to the farthest separation of the two classes. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized [3].

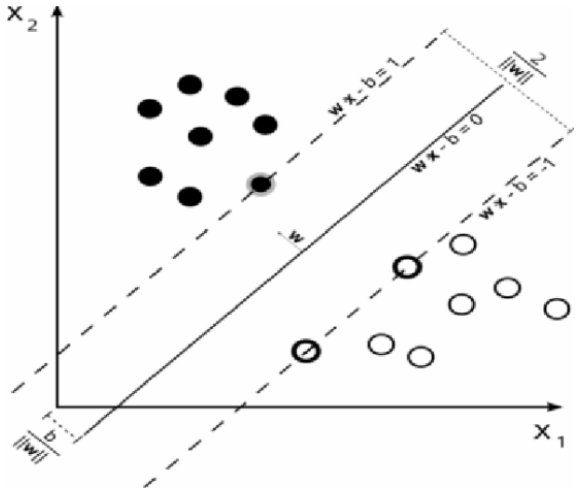

Fig. 2: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors *[4]*

B. K-nearest neighbors（KNN）

KNN is one of the most important supervised machine learning algorithms, often used for classification problems. Since the KNN algorithm is lazy learning, there is no training phase. To classify, this algorithm determines the class with the k closest classes as the class of the new data. Determining the value of k is important here. The k value may take a different optimum value for each problem. It uses the Euclidean distance calculation to find the distance of the point to be predicted to other points [5].

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

In our study, we used the optimal value of k as 7. While calculating the k value, we took the k value from 1 to 25 and examined the accuracy score values. We found that the k value is more successful when the accuracy score is 7.

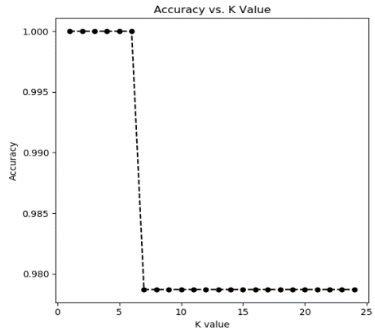

Fig. 3: Best fit k value

C. *Keras*

Keras is a deep learning library. Keras enables fast creation and training of deep learning models. When building deep learning models in Keras, the second layer understands the output of the first layer, so we don't need to specify the output each time in the input of the next layer. Keras is distributed under the permissive MIT license, which means it can be freely used in commercial projects. It's compatible with any version of Python from 2.7 to 3.6 (as of mid-2017). In the application to be made with Keras, we need to prepare the data, define the model and give the data as input to the layers, determine the activation functions and train the model with the fit function. Keras models are of two types: Sequential and Functional API. The most widely used model is the Sequential model. A Sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor [7]. Keras provides industry-strength performance and scalability: it is used by organizations and companies including NASA, YouTube, or Waymo [8].

D. *StratifiedKfold Cross-validation*

Once our model is trained, we move on to the validation phase. Sometimes, when the size of our data is small, we may have to test our model with a small number of data. When the validation process is with the same data, it is not very healthy. When we test with different data, we get different scores. StratifiedKfold cross validation is used for unbalanced data sets. It ensures that both groups of data are included in the test data. In the standard k-fold method, there is a possibility that both types of data may not be present. It divides the data into k parts and performs data training with the k-1 part and validation with the rest.

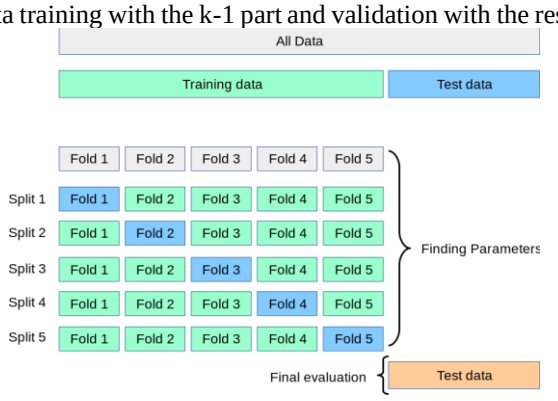

Fig. 4: Cross-validation *[9]*

Validation score is the average of the validation score of k parts. When building the models, the data is divided as 25% test and 75% training data. When the accuracy of the model is calculated with this split, the success rate may be higher than normal due to overfitting. StratifiedKfold cross validation prevents this situation. Since it validates with different test data, we get different validation scores. By averaging the validation scores, a more realistic score is obtained.

IV. Experimental Results

Raman spectra of four different paper types were obtained by min-max normalization and baseline correction and are shown in Fig. 5. The spectra show Brazilian A4, standard A4 and also diploma papers with different thicknesses. The differences in the types of these papers can be clearly recognized by the peaks in the Raman spectra. In the 1300 - 1500 cm-1 region, especially diploma papers and A4 papers show similar characteristics and can be distinguished from each other. The peak at approximately 1100 cm-1 raman shift value present in all papers comes from the structure of cellulose [6].

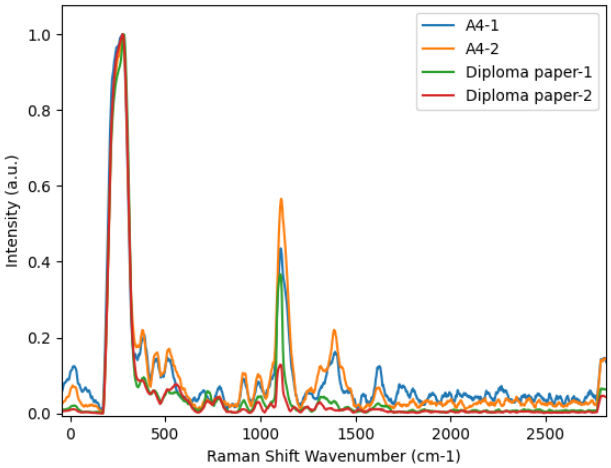

Fig. 5: Different Paper Raman Spectrum

Fig. 6, Old Pen 1 and Old Pen 2 are new samples taken from a Schneider Xtra 8053 pilot pen that was opened and used about 5 years ago and they differ in the way they are written. Old Pen-1 is the spectrum of the sample that is written thinly, without pressing down while writing. The Old Pen 2 is the spectrum of the data drawn thicker with pressure. New Pen-1 and New Pen 2 are samples taken using a newly purchased Schneider Xtra 8053 pilot pen of the same brand. New Pen 1 is the spectrum of the samples written thinner and New Pen 2 is the spectrum of the samples written thicker. Fig. 6 shows that there is no spectral difference between the old and new pen. Old Data 1 spectrum is the spectrum taken from A4 paper drawn about 5 years ago. Regardless of whether the pen is old or new, it is observed to be different from the old written data.

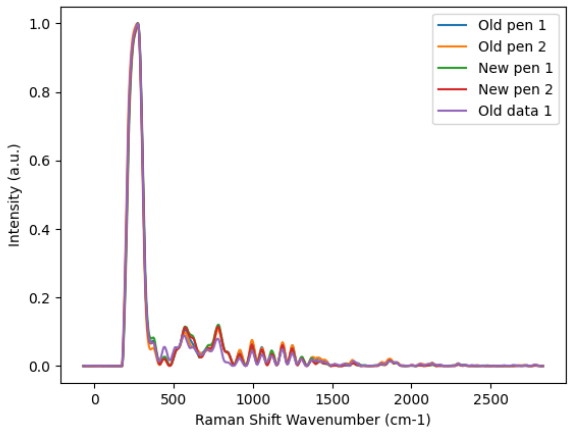

Fig. 6: Effects of Pen Thickness and Condition on Raman Spectrum

Fig. 7 shows the Raman Spectra of three samples written five years ago and three newly written samples. When the Raman Spectra of the old and new pencil samples are analyzed, no dominant change in the peaks is observed, but when the 300 - 1000 cm-1 region is analyzed in Fig. 8, when the 300 - 1000 cm-1 region is examined, the data of the old and newly written samples are separated within themselves, especially when examined in the 400-500 cm-1, 550-600 cm-1, 750-800 cm-1 regions. This separation is due to the changes in the molecular structure of the pen ink in the air environment over time. These changes are sufficient to solve the classification problem and the changes can be detected as features by machine learning algorithms.

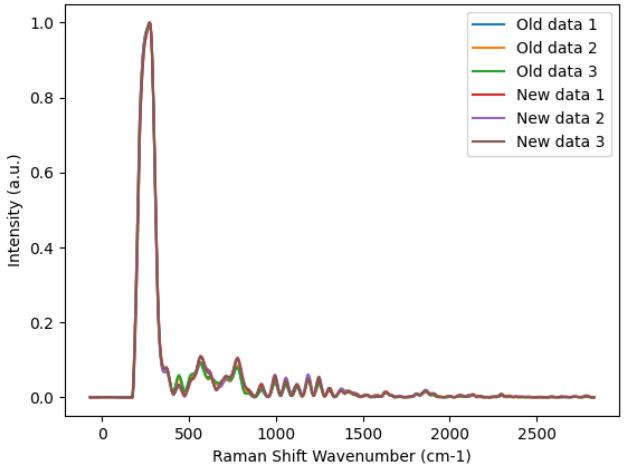

Fig 7: Effects Of Pencil Aging On Raman Spectrum

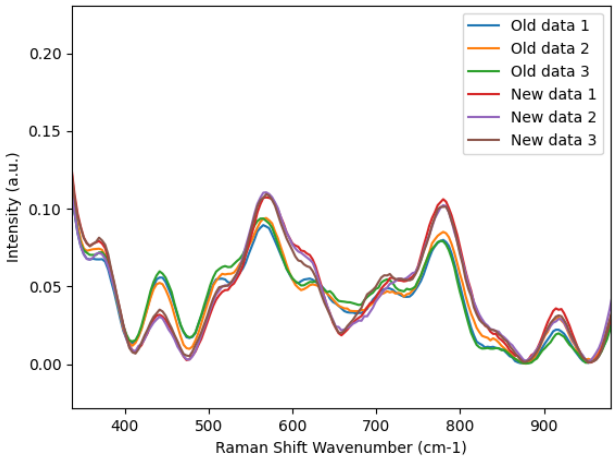

Fig. 8 : Effects Of Pencil Aging On Raman Spectrum (300-100 cm^{-1})

Python version 3.7.6 was used in our study. Spyder 4.0.1 IDE was used for code development environment. For the classification problem, we worked with SVM and KNN algorithms using Python Scikit-learn library. A neural network was developed using Tenserflow Keras library.

Data were collected from different people (men and women) at different times. 155 data were used in our study. Spectra of the collected data were measured using a Raman QE Pro High spectrometer. Min-max normalization was performed to fix the peaks of the spectrum data to 1. In our study, binary classification was performed to predict whether it was old or new writing. SVM, K-nearest neighbors (KNN) and Neural Network are the most common methods used for classification problems.

Data Type	Data Count
Old Data	84
New Data	71

Fig. 9: Number of data used

Support Vector Machine (SVM) and KNN are machine learning methods frequently used in classification problems. In SVM and KNN algorithms, 70% of the data is used for training and 30% for testing. In the KNN algorithm, the k value is taken as 7. It is concluded that the accuracy of both algorithms is 100%. When developing the Keras Sequential model, the hidden layers use the Rectified Linear Unit (ReLU) activation function. The sigmoid activation function is used in the output layer. ReLU activation function outputs between [0, +∞]. Since ReLU takes numbers less than 0 as 0, it does not produce negative outputs and is only active for positive outputs. This affects the performance positively. Adam optimization function is used in our model. The input data size of our model is 1037. Our model consists of 3 Dense. The output activation function is Sigmoid. Sigmoid takes values between [0,1]. It is a non-linear function. StratifiedKFold cross validation divided into 5 parts. The success rate of the model was 98.71%. KNN and SVM had a higher accuracy score than Neural network.

V. CONCLUSION AND FUTURE WORK

Although the classification of pen ink aging using Raman spectra has been successful, it does not provide information about the molecular chemical structure of the ink due to the predominance of fluorescence signals in this region. Solutions can be evaluated by taking Raman spectra with a different wavelength laser source in order to take into account the cases where the writing is done with another pen. Since our current system is compatible with the 785nm laser source, we have not yet been able to perform this study. In addition, the evaluation of the samples written with FTIR (Fourier Transform Infrared Spectrospy) spectra is planned as our further studies. It is planned to collect data on different types of paper and conduct a similar study. Classification with CNN algorithm is also planned.

ACKNOWLEDGMENT

The data measurement phase of this study was completed using the devices and auxiliary instruments in TÜBİTAK BİLGEM Optics Laboratory. We would like to thank TÜBİTAK BİLGEM for their support.

REFERENCES

[1] İ. Birincioğlu ve E. Özkara, «Adli Belge İncelemelerinde Bilinmeyenler,Örneklerle Yazı ve İmza Analizi ile Islak İmza Kavramı,» *TBB Dergisi*, pp. 403-433, 2010.

[2] İ. Çakır ve H. Aslıyüksek, «Instruments and Methods in Forensic Document,» *Arşiv Dünyası,* pp. 16-17, 2014.

[3] A. T. Harris, A. Lungari, C. J. Needham, S. L. Smith, M. A. Lones, S. E. Fisher, X. B. Yang, N. Cooper, J. Kirkham, D. A. Smith, D. P. Martin-Hirsch ve A. S. High, «Potential for Raman spectroscopy to provide cancer screening,» *Head & Neck Oncology*, cilt 1, no. 1, p. 34, 2009.

[4] . A. Emin, A. Hushur ve T. Mamtimin, «Raman study of mixed solutions of methanol and ethanol,» *AIP Advances*, cilt 10, no. 6, p. 2020, 2020.

[5] R. Pueyo, M. Soneira ve S. Ruiz-moreno, «Morphology-Based Automated Baseline Removal for Raman Spectra of Artistic Pigments,» *Applied spectroscopy*, cilt 64, no. 10.1366/000370210791414281, pp. 595-600, 2010.

[6] U. Agarwal, «Analysis of Cellulose and Lignocellulose Materials by Raman Spectroscopy: A Review of the Current Status,» *Molecules*, cilt 1659, no. 10.3390/molecules24091659, p. 24, 2019.

[7] Y. Lin ve J. Wang, «Research on text classification based on SVM-KNN,» %1 içinde *2014 IEEE 5th International Conference on Software Engineering and Service Science*, Beijing, China, 2014.

[8] Y. Lin ve J. Wang, Artists, *Maximum-margin hyperplane and margins.* [Art]. IEEE.

[9] M. A. Lusiandro, S. M. Nasution ve C. Setianingsih, «Implementation of the Advanced Traffic Management System using k-Nearest Neighbor Algorithm,» %1 içinde *2020 International Conference on Information Technology Systems and Innovation*, Bandung, Indonesia, 2020.

[10] F. Chollet, DEEP LEARNING with PYTHON, United States of America: Manning Publications Co, 2018.

[11] F. Chollet, «The Sequential model,» 12 04 2020. [Çevrimiçi]. Available: https://keras.io/guides/sequential_model/.

[12] «About Keras,» 2015. [Çevrimiçi]. Available: https://keras.io/.

[13] «Cross-validation: evaluating estimator performance,» 19 10 2011. [Çevrimiçi]. Available: https://scikit-learn.org/.