DOI: http://doi.org/10.4038/sljp.v23i1.8110

INSTITUTE OF PHYSICS – SRI LANKA

Research Article

A quantitative approach to gemstone identification using Raman spectroscopy combined with machine learning

W. I. W. Y. Boteju[1], M. K. A. S. Gunarathna[1], T. L. M. D. Fonseka[1], H. N. S. Peiris[2], L. A. R. Silva[2], N. N. S. S. Jayathilaka[2], L. A. G. D. Jayasekara[1], P. N. Perera[3], H. H. E. Jayaweera[1], M. S. Gunewardene[1], S. Jayawardhana[2]*

[1]*Department of Physics, University of Colombo, Colombo 03, Sri Lanka*
[2]*Department of Physics, University of Sri Jayewardenepura, Gangodawila, Nugegoda, Sri Lanka*
[3]*Zymergen Inc. Emeryville, CA, USA*

Abstract

Raman spectroscopy is an ideal technique for gemstones identification due to its non-destructive nature, rapid detection, no sample preparation, and ability to analyze interior compositions. Notwithstanding the benefits, most routine gemstone analysis requires complementary techniques to verify the accuracy due to difficulties in matching Raman spectra against a known database, while ensuring high sensitivity, specificity, and accuracy. This work presents a technique where computational methods are used to accurately identify gemstones for routine operations. The acquired Raman spectroscopic data is pre-processed using baseline subtraction and signal smoothing for optimal signal extraction and then cross correlated with a verified database of spectra. The correlation coefficient results are then clustered using a K-means algorithm to distinguish the gemstone families. The locally sourced unknown gemstone was found to belong to the quartz family with a sensitivity of 100%, specificity of 98% and accuracy of 98%. A second technique was also introduced by considering both cross correlation and overlap between the area under the curves of matched spectra. Both methods converged on the same conclusion and was backed up by three common cluster validation indices thereby assuring the validity of the identification. The technique was further validated to be used with other gemstone families such as beryl, diamond, and corundum.

Keywords: Raman spectroscopy, Machine learning, gemmology, clustering, k-means, correlation coefficient, cross-correlation.

∗ Corresponding author: **sasanijay@sci.sjp.ac.lk**

https://orcid.org/0000-0003-1532-393X

1. INTRODUCTION

Gemstones are crystalline minerals formed over time under favorable geological conditions[1]. Their rarity, durability and aesthetic beauty make them highly valuable as a commodity. Gemstones can take different shades of color, clarity, luster, and sheen which makes it important for a jeweler to identify them accurately in order to both cut and to certify its authenticity. Furthermore, with the development of technology, there is an increasing number of laboratory-made synthetic gemstones as well as heat-treated and modified gemstones. Therefore, the accurate identification and classification in terms of mineral species, purity, and country of origin are of utmost importance[2].

The identification of smaller gemstones was traditionally carried out with the aid of a field expert equipped with a spectroscope. However, this is insufficient for valuable gemstones that require certification. Most gemological institutes resort to multiple complementary techniques such as infrared, UV-Visible spectrometry, refractometry, X-ray diffraction and a basic version of Raman spectrometry. Although there are a variety of advanced techniques that can be used for the purpose which includes, chemical microanalysis technique such as LA-ICP-MS (Laser Ablation–Inductively Coupled Plasma–Mass Spectrometry), LIBS (Laser-Induced Breakdown Spectroscopy); SIMS (Secondary Ion Mass Spectrometry), photoluminescence spectroscopy; cathodoluminescence spectroscopy, real-time fluorescence imaging, x-ray imaging; energy-dispersive x-ray fluorescence (EDXRF), x-ray radiography & tomography and Focused Ion Beam (FIB), their destructive nature makes them less-than ideal for routine gem identification[3].

Nevertheless, Raman spectrometry is a non-destructive technique which requires no sample preparation or modification. Considering the low cost of operation, short measurement times and its ability to characterize micro-crystalline features such as inclusions make it an ideal candidate for gem identification[2,3,4]. A Raman spectrum displays clearly defined intensity peaks as a function of Raman shifted wavenumbers. Since these peaks correspond to vibrational state energies of the molecules present in the sample, it provides a unique spectral ‘fingerprint’ of the chemical compound thereby enabling identification. For a routine sampling method in material identification, the process requires the acquired Raman spectra to be matched against a known database of chemical compounds. Databases for material categories such as polymers, pharmaceutical drugs, narcotics and biomarkers are already commercially available. Therefore, such test

spectra can be directly compared against these databases using simple statistical techniques. However, unlike the above compounds which consist of a specific chemical makeup that gives rise to precise Raman peak characteristics, minerals have the potential to vary in their chemical composition. This is a result of the multitude of physical conditions such as temperature, pressure, soil composition, and humidity that is present throughout the formation of the stones[5]. As a result, although gemstones from the same gem family are usually built on similar ideal chemistry, subtle variations in the acquired Raman spectra are common[6]. This makes it difficult to utilize simple statistical techniques for spectral matching against a known database for gemstone identification. On the other hand, subtle differences in spectra can provide additional information on the specific gem type, crystal orientation, origin and quality of a particular stone[1]. Expert knowledge is required to extract fine details from such a spectrum, which naturally leads to machine learning techniques.

Machine learning has been a recent area of interest in gemstone identification. Deep learning models can be easily used to identify and classify gemstones by training the algorithm to extract information from a set of gemstones images[7]. The results of such an approach rely heavily on image quality factors such as the nature of illumination used, camera distance, and quality of the image sensor. Therefore, such an approach requires the acquisition of a comprehensive set of laboratory-controlled gemstone images to be used as the training dataset.

A more comprehensive method had been adapted by Pena et al, to categorise and grade emerald stones[8]. Here, both unsupervised clustering and supervised classification methods have been used to determine the quality and therefore the pricing of emerald stones. Despite the potential, both are ‘visual’ methods that are yet to be approved to be used as a legitimate method for gemstone certification.

Contrastingly, Raman spectroscopy is already an established technique that has been approved for its use in gemmology laboratories for certification. Nevertheless, its use has been limited due to the difficulties of spectral matching and feature extraction needed for routine analysis.

Diez-Pastor et al, used machine learning to investigate the Raman spectra of mineral Variscite obtained from an ancient mining complex in Gava, Europe[9]. They compared a

number of classification algorithms to establish a relationship between the sample and its originating mine and depth. Although limited in its application to a single mineral from a single origin, this provides interesting details on how machine learning can be used to extract detailed information that was previously unavailable.

This paper expands on existing literature by putting forward a computational technique encompassing both statistical techniques for pre-processing data and machine learning techniques for the identification of gemstones for a wide range of families. This builds up to a single powerful technique that can be utilized by non-experts for routine and rapid gemstone characterization. Routine analysis can potentially be handled by machine learning techniques with a comprehensive database of Raman spectra[10]. One of the most widely used databases originates from the RRUFF Project at the University of Arizona (Tucson, Arizona), which is an open database that lists a considerable number of minerals with their spectral data and supplementary information. This enables researchers to extract spectral data to custom build databases with spectral searching functionality. Spectral matching methods such as Pearson's correlation coefficient[11], cross- correlation[12] can be used to identify the relevant gemstone with the information from the database. With the aid of machine learning techniques, it enables the identification and categorization of gemstones into their gemstone families more efficiently and accurately. Hence, two methods are presented here for lining up spectral peaks of an experimental spectrum with a standard, quantifying the degree of agreement and identifying the gem family with the correlation coefficient and cross- correlation matching.

2. METHODOLOGY

An uncut and unpolished gemstone sourced from a Sri Lankan gem deposit was selected for the analysis. The sample had a purple hue and appeared to have a reasonable degree of clarity.

A custom-built Raman spectrometer was used for the acquisition of Raman spectra. The setup consisted of a 532 nm DPSS laser (Cobolt Samba) for excitation, a x50 objective (Mitutoyo plan apocromat) was used in the backscattered geometry to collect the scattered signal which was filtered through a Raman edge filter with a steep 537.2 nm cutoff (Semrock, RazorEdge) and guided through to the Andor Kymera 328iA spectrograph for analysis. A 600 lines per mm grating was used with a thermo-electrically cooled electron-

multiplying charge coupled device (EMCCD) for detection. The spectrum of the sample gemstone was captured through this setup using the Solis Acquisition software. In the acquisition of Raman spectra 10 measurements were taken where the sample was rotated after each measurement. Since crystalline minerals could display subtle differences in peak intensities due to orientation, this approach ensures such effects to be averaged out in the analysis. Nevertheless, the samples analyzed in this work did not display any orientation-specific spectral characteristics, therefore the differences between the averaged spectrum and specific spectra were negligible.

The spectral analysis was conducted through a self-written program based on Python 3.8. The program was executed on a personal computer AMD Ryzen 5 5600H, with Clock speed 3.3 GHz, and Main memory 16 GB. Modified asymmetric least squares method was used as the baseline removal technique[13]. Spectra were denoised by using a Savtizky-Golay filter with a window size of 15 and a 4-degree polynomial. Furthermore, the baseline was refined by manual inspection whenever required.

A pool of 115 standard spectra of gemstones was obtained from the RRUFF database[14]. Thus, gathered standard spectra were down sampled to match indices of spectra to match the nearest key. Experimental spectrum and the standard spectra were normalized to the corresponding maximum peak height. The experimental spectrum was padded with two null data points on either side, allowing to absorb a tolerance of a ±4 Raman shifts. Each standard spectrum under the investigation was slid through the padded experimental spectrum considering the cross-correlation. The experimental data set was shifted to the position with the maximum cross-correlation along the Raman shift (see Figure 1).

The correlation coefficients were determined. The machine learning algorithm, K-Means clustering was performed on the 1D array of correlation coefficient with a consistently automated threshold (method 1). The Clustering Validation indices (CVI) Silhouette coefficient, Davies–Bouldin index, and Calinski–Harabasz index were calculated[15]. Ultimately, the gemstone family was determined by means of visual inspection, correlation coefficient, and K-Means clustering.

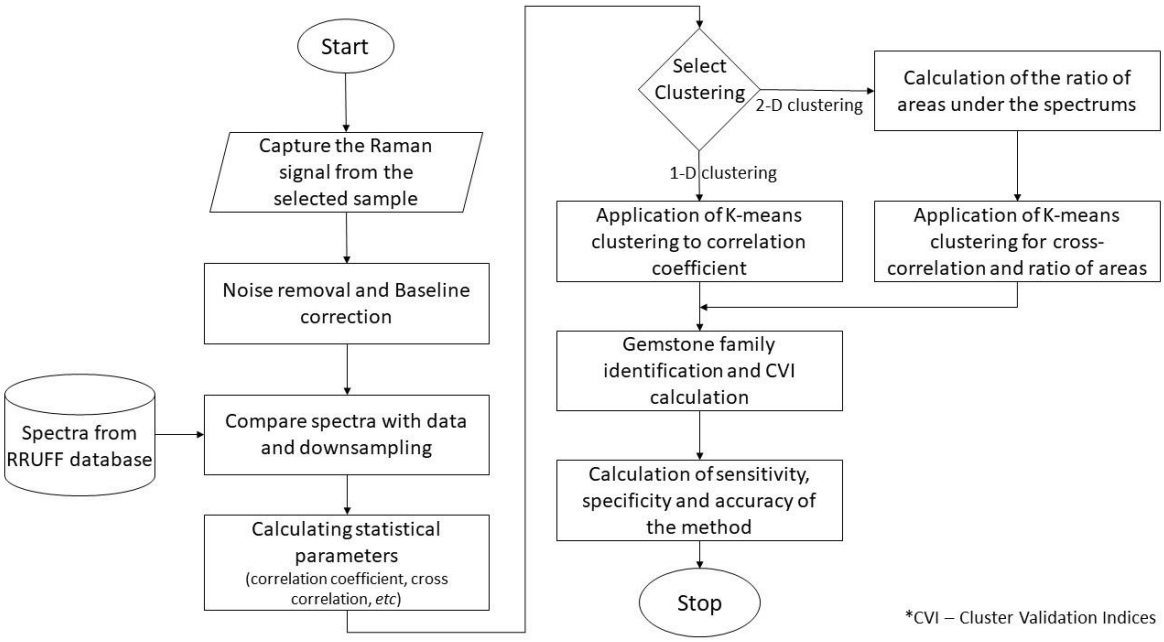

Figure 1: Block diagram of the methodology

An optional method of determining the gem family of the unknown gemstone was introduced. The ratio of the area under the standard graph to the area under the experimental graph was calculated. Two-dimensional K-Means clustering was implemented on the data set which consists of the calculated ratio and the cross-correlation[16]. Previously mentioned CVIs were recalculated. 2D K-Means clustering was used to verify the previous classification of the gem family (method 2).

The reliability of the proposed techniques was individually assessed by using the parameters Sensitivity, Specificity, and Accuracy[17,18]

3. RESULTS AND DISCUSSION

This section consists of three segments which are dedicated to the results of spectral preprocessing, identification of the gemstone family from K-means clustering method 1 and correlation coefficient, and validation of the gemstone family using K-means clustering method 2, cross-correlation and the ratio of area under the graph.

3.1. Spectral preprocessing and downsampling

The Raman spectrum of the unknown gemstone ranges from 0 to 3500 cm^{-1} with prominent Raman peaks at 124.6 cm^{-1}, 200.71 cm^{-1}, 351.33 cm^{-1}, 389.88 cm^{-1}, 458.86 cm^{-1} and 1156.06 cm^{-1}. The baseline-corrected graph of the unknown gemstone was obtained by

applying a Savitzky - Golay filter to reduce the high-frequency noise component. The window size and degree of the polynomial of this filter were specifically adjusted to retain the characteristics of the raw spectra. As the standard approach, the baseline-corrected graph was limited from 0 to 1800 cm^{-1} since this is considered as the fingerprint region of Raman spectroscopy. Manual refinement of the baseline was important in certain cases to retain the proportionate height of the original intensity peaks.

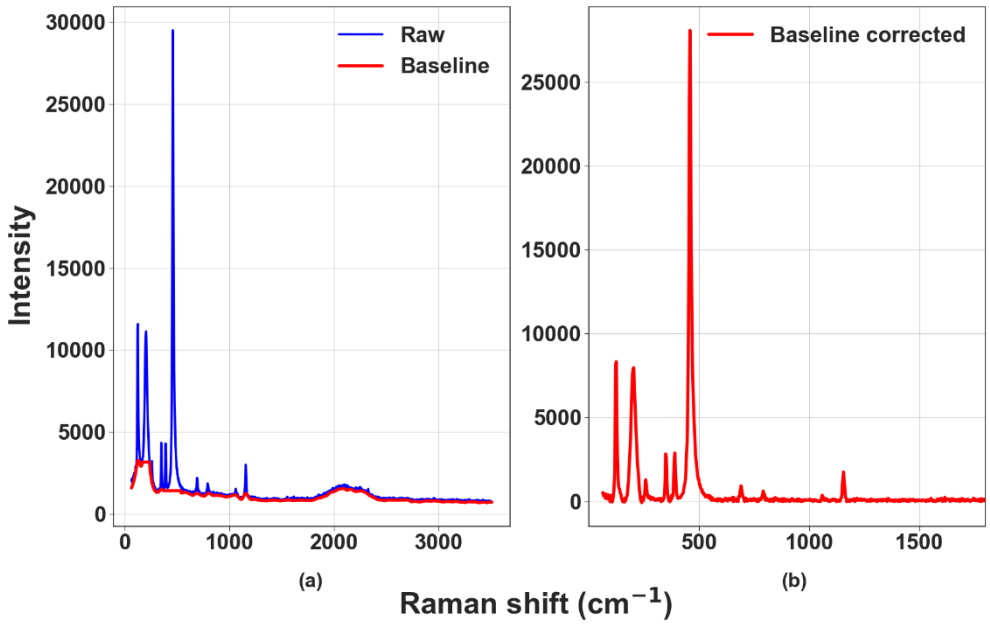

Figure 2: The Raman spectra of the unknown gemstone where (a) illustrates the raw spectrum along with the fitted baseline, and (b) the baseline corrected spectrum.

The standard datasets extracted from the RRUFF database, and the experimental dataset were resized and downsampled so that they have approximately equal starting points, ending points and an equal number of data points within the spectral region.

The datasets were downsampled with a tolerance of 5, this allowed the datasets to map a specific value from the lower resolution dataset to the higher resolution dataset within a range of ±5 cm^{-1}. The lengths of the resized datasets were decided using the starting and ending points of both graphs in comparison (figure 3).

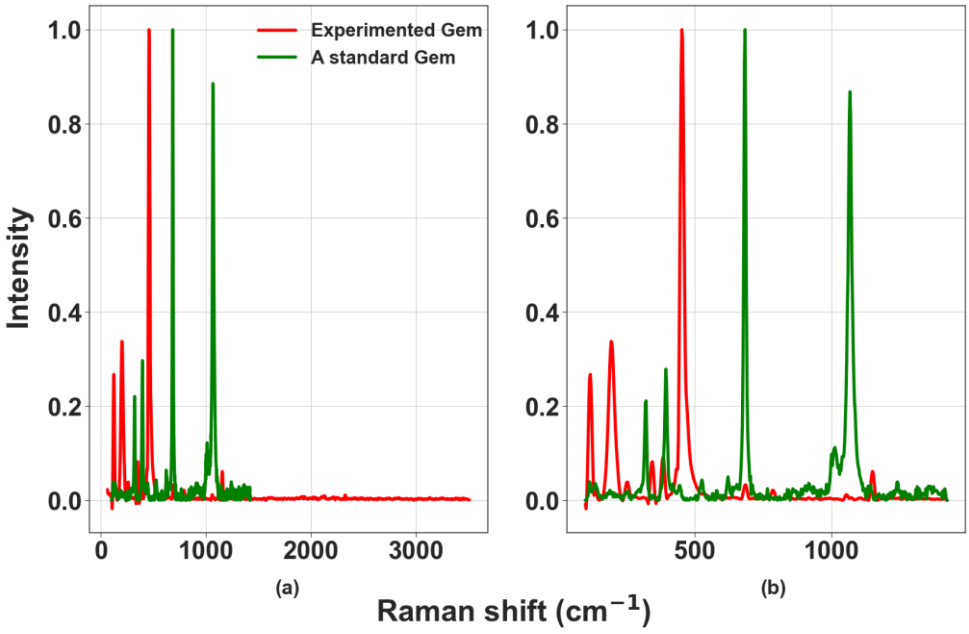

Figure 3: The spectrum of the unknown experimental gemstone overlaid with (a) The original standard gemstone spectrum and (b) Downsampled and resized standard gemstone spectrum.

3.2. Identifying unknown gem families using K-means algorithm

The correlation coefficient values obtained through cross-correlation of the experimental spectra against each spectrum taken from the database were then used as inputs in a K-means clustering algorithm. This unsupervised machine learning algorithm aims to cluster data points such that it minimizes the Euclidean sum-of-squares distances to a given centroid.

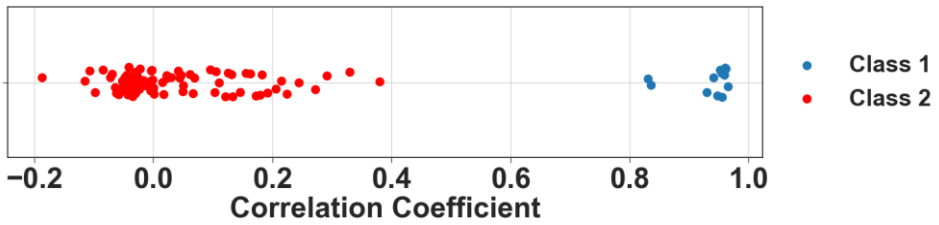

Figure 4: Graph of 1D K-Means clustering illustrating two clear classes of clusters. Note the vertical spread of data points in illustrate the density of the distribution.

Correlation coefficient values vary within the range [-1, 1] with 1 being a strong positive correlation, -1 being a strong negative correlation and 0 indicating no correlation. According to figure 4, the densest area in terms of datapoints is closer to 0. This depicts the non-existence of a relationship with most of the standard datasets. Another slightly dense area is visible closer to 1. This indicates a small group of standard datasets which have a

strong relationship with the dataset of the unknown gemstone. Hence, the family of the pool of standard datasets clustered as class 1 can be considered as the potential candidate for the family of the unknown gemstone. This particular group of spectra obtained from the database is attributed to the quartz (silicate) family of gemstones. A slightly separated sub-cluster of data points can also be seen in class 1, which has a marginally lower correlation coefficient just above 0.8. These correspond to a couple of spectra associated with a rare gemstone called Brannockite.

Figure 5 shows the overlay of spectra from class 1 as identified in Figure 4. All quartz graphs show a similar variation in intensities along with the graph of the unknown experimental gemstone at the bottom. However, upon visual inspection it can be seen that the Brannockite graphs show a slightly different intensity variation pattern. This is in agreement with the cluster analysis results and hence, the Brannockite gemstone can be declared as a mismatch with the graph of the unknown experimental gemstone. Therefore, the two Brannockite standard samples are considered as false positives. Altogether, 10 true positive results, 103 true-negative results, two false-positive and zero false-negative results were obtained. The standard parameters of the algorithm specific to this unknown gemstone were obtained.

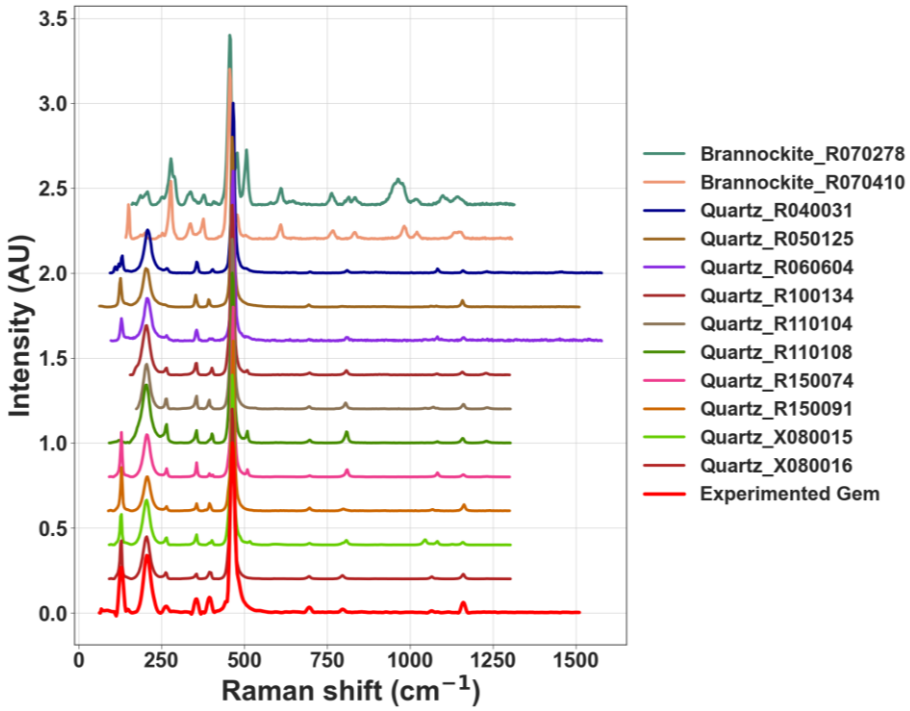

Figure 5: Graphs of standard datasets in class 1 with the dataset of the unknown experimental gemstone at the bottom.

Table 1: Table of the standard parameters. Sensitivity states the ability to detect a true positive. The upper bound of this parameter is 1. Hence, the algorithm has a greater chance to detect a true positive accurately. Specificity is the ability to detect a true negative. The accuracy of the algorithm for the unknown experimental gemstone is 98.26%

Parameter	Value
Sensitivity	100.00%
Specificity	98.10%
Accuracy	98.26%

The result from the above method was verified using 2D K-Means clustering of cross correlation and the ratio of the area under the standard graph to the area under the experimental graph. The ideal cluster should have a value closer to 1 for the ratio of the areas and a maximum value for the cross-correlation. Therefore, the class 4 cluster can be considered as the class with the potential unknown gemstone family.

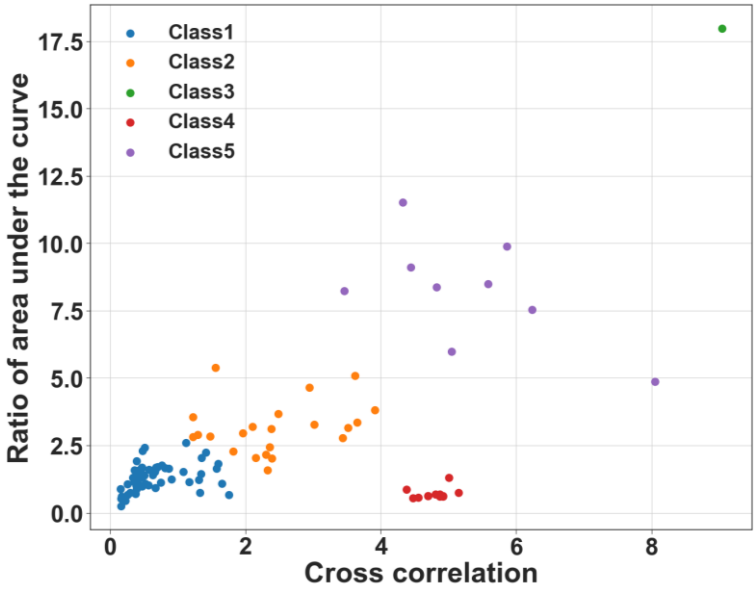

Figure 6: Graph of 2D K-Means clustering. The datasets are clustered considering the cross correlation and the ratio of the area under the standard graphs and the experimental graph of the unknown gemstone. The ideal cluster should have a value closer to 1 along the y-axis with a maximum value for cross-correlation.

Note the outlier in class 3. It has a stronger cross correlation than class 4. As can be seen in figure 7, the common area under the curves is almost equal to the area under the experimented gem. Therefore, a higher cross correlation is obtained. However, this is a misleading result. This falsified outcome is omitted by the 2D K-Means clustering and can be verified by visual inspection. Hence, class 3 was omitted from the possible candidates for the unknown gem.

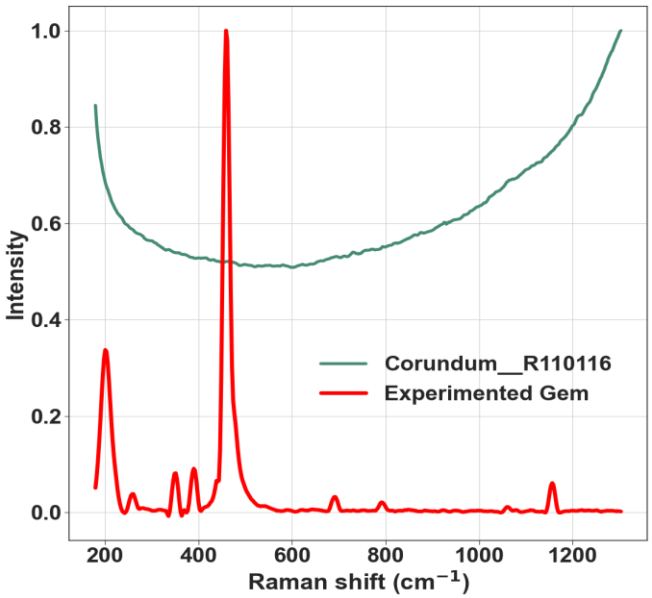

Figure 7: Graph corresponding to the outlier in class3 of figure 6, with the Raman spectrum of the experimented unknown gemstone plotted against the outlier corundum spectrum extracted from the database.

The obtained results from both clustering methods have the same set of candidates. Hence, the result from 1st K-Means clustering method was verified to be accurate using the 2nd K-Means clustering method. Validity of the machine learning algorithms was checked and compared using three Cluster Validation Indices (CVI) as summarized in Table 2.

Table 2: Table of CVIs. The Silhouette coefficient provides a value closer to 1 which indicates classes are densely clustered. Davies–Bouldin index shows a minimal value which is another indication of strong clustering. Furthermore, the larger the Calinski–Harabasz index as shown in the below table, the stronger the clustering.

Cluster Validation Indices	**K-Means method 1**	**K-Means method 2**	**Remarks**
Silhouette coefficient	0.88	0.60	Closer to 1 is better
Davies–Bouldin index	0.12	0.49	Closer to 0 is better
Calinski–Harabasz index	943.16	249.35	The higher the number the better clustering

As mentioned above, the unknown gemstone family was determined to be quartz using spectral analysis incorporated with the machine learning technique of K-Means clustering. The final verification of the result can be testified by visual inspection. Moreover, the sampled gemstone was assessed as an Amethyst (Quartz family) by a field expert in gemmology thereby further verifying the result. The validity of the machine learning algorithm was justified by means of CVIs.

3.3. Validation of the algorithm for different gem families

The algorithm was further validated for the gem families of Beryl, Garnet, Corundum and Diamond, the spectra of which were extracted from the database. Raman spectra of a sample verified through a complementary technique such as X-ray diffraction (XRD) was used as the benchmark to compare the validity of its match with other gemstones from the same family. The family of a known Beryl sample ($Be_3Al_2Si_6O_{18}$) was verified to be Beryl with 18 true positive values from a pool of 18 Beryl datasets. A known diamond sample was verified to be a diamond with 15 true positive values from a pool of 15 diamond datasets. The family of a known Corundum sample was verified to be Corundum with 15 true positive values from a pool of 17 Corundum datasets. The family of a known Garnet sample was verified to be Garnet with 02 true positive values from a pool of 07 Garnet datasets.

The sensitivity, specificity and accuracy of the technique for each of the gemstone families is quantified in table 3. It should be noted that the Beryl and Diamond families show perfect matches since the ideal chemistry of the gemstones within these families show almost no variation as can be seen in Figure 8. In contrast, due to the varying levels of impurities and inclusions present across the members of the corundum family, which results in some variation across their spectral characteristics, the sensitivity and accuracy of the technique is marginally lowered. This is further aggravated for the Garnet family since Garnets are complex minerals with a general chemical formula of $X_3Y_2(SiO_4)_3$ were X could take any one of Ca, Fe^{2+}, M, Mn^{2+} and Y could take $Al, Cr, Fe^{3+}, Mn^{3+}, Si, Ti, V\ or\ Zr$. Nevertheless, subject to the availability of a comprehensive database of Raman spectra that includes a reasonable sample size of spectra from each chemical compound, the spectral matching techniques proposed here can be accurately utilized.

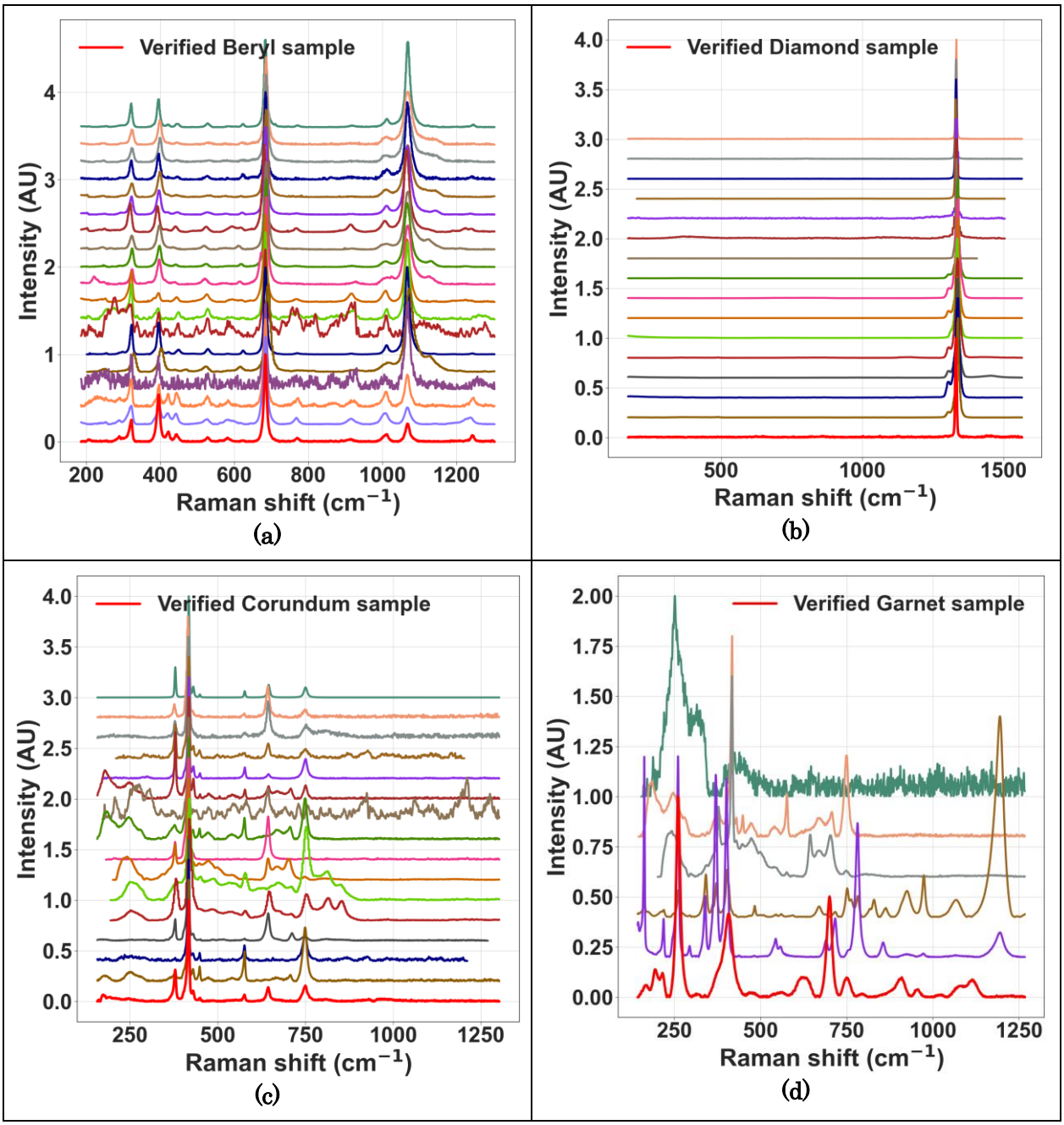

Figure 8: Overlay of Raman spectra for the gemstone families (a) Beryl (b) Diamond (c) Corundum and (d) Garnet.

Table 3: Table of performance metric for known gemstones.

	Sensitivity	Specificity	Accuracy
Beryl	100.00%	100.00%	100.00%
Diamond	100.00%	100.00%	100.00%
Corundum	88.24%	100.00%	98.26%
Garnet	28.57%	97.22%	93.04%

4. Conclusion

The proposed Raman spectral matching techniques provides a highly sensitive and accurate tool to carry out gemstone identification in real-time. The two techniques that were investigated converge on the same conclusion thereby assuring the validity of the identification. Further, three common CVIs were also computed to verify the use of the K-means clustering machine learning techniques.

The technique can be deployed with a high level of accuracy (>98%) for the gemstone families of Beryl, Diamond and Corundum. The accuracy can be further improved with the expansion of the Raman spectral database of minerals.

The computational time for the automated spectral matching functionality from beginning to end consumed 3.189 seconds thereby enabling rapid identification of a gemstone variety.

Due to its non-destructive nature, ability to penetrate deep into a sample to analyze inclusions and high throughput, Raman spectroscopy coupled with machine learning techniques paves the way for an all-encompassing technique for the routine identification of gemstones with a reduced requirement on expert knowledge.

References

1. Tariwonga, Y. *et al.*, (2020). X-ray induced luminescence, optical, compositional and structural investigations of natural and imitation rubies: Identification technique. *Radiation Physics and Chemistry*.177, DOI: https://doi.org/10.1016/j.radphyschem.2020.109089

2. Bersani, D., Lottici, P. (2010). Applications of Raman spectroscopy to gemology. *Analytical and Bioanalytical Chemistry*. 397, pp.2631-2646. DOI: https://doi.org/10.1007/s00216-010-3700-1

3. Breeding, C. M. (2010). Developments in Gemstone Analysis techniques and Instrumentation During the 2000s. *Gems and Gemology*. 46, DOI: http://doi.org/10.5741/GEMS.46.3.241 .

4. Raneri, S. *et al.*, (2020). Non-destructive spectroscopic methods for gem analysis: a short review. *Metrology for Archaeology and Cultural Heritage*.

5. Groat, L. (2018), Scientific Study of Colored Gem Deposits and Modern Fingerprinting Methods, *Gems & Gemology*, Gemological Institute of America, pp.277-278

6. Using Raman Spectroscopy for Gemstone Analysis, viewed 15th November 2021 https://www.labcompare.com/10-Featured-Articles/338456-Using-Raman-Spectroscopy-for-Gemstone-Analysis/

7. Gemstone Classification using Deep Learning, viewed 21st November 2021 https://medium.com/ai-techsystems/gemstone-classification-using-deep-learning-46ea270a4c03

8. Pena, F. B. *et al.,* (2021). Machine learning applied to emerald gemstone grading: framework proposal and creation of a public dataset. *Pattern Analysis and Applications*, 25, pp.241-251, DOI: https://doi.org/10.1007/s10044-021-01041-4

9. Díez-Pastor, J. F. *et al.*,(2018). Machine learning algorithms applied to Raman spectra for the identification of variscite originating from the mining complex of Gavà, *Journal of Raman Spectroscopy*, 51, pp.1563-1574, DOI: https://doi.org/10.1002/jrs.5509

10. Lowry, S. *et al.,* (2009). The Use of a Raman Spectral Database of Minerals for the Rapid Verification of Semiprecious Gemstones. *Spectroscopy (Santa Monica)*. 24.

11. Samuel, A. Z. *et al.,* (2021). On Selecting a Suitable Spectral Matching Method for Automated Analytical Applications of Raman Spectroscopy. *ACS Omega*. 6, pp.2060-2065. DOI: https://doi.org/10.1021/acsomega.0c05041.

12. Hui, C. *et al.,* (2018). Eliminating Non-linear Raman Shift Displacement Between Spectrometers via Moving Window Fast Fourier Transform Cross-Correlation. *Frontiers in Chemistry*. 6, pp.515. DOI: https://doi.org/10.3389/fchem.2018.00515.

13. Oller-Moreno, S. (2014). Adaptive Asymmetric Least Squares baseline estimation for analytical instruments. *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, pp.1-5. DOI: http://doi.org/10.1109/SSD.2014.6808837.

14. Database of Raman spectroscopy, X-ray diffraction and chemistry of minerals, viewed 20th November 2021 https://rruff.info/

15. Arbelaitz, O. *et al.,* (2013) An extensive comparative study of cluster validity indices. *Pattern Recognition*. 46, pp.243-256. DOI: http://doi.org/10.1016/j.patcog.2012.07.021.

16. Boyd, D. W. (2001). Systems Analysis and Modeling, *Academic Press*, pp.211-227, DOI: https://doi.org/10.1016/B978-012121851-5/50008-3

17. Robert, T. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Liabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*. 5, p.307, DOI: https://doi.org/10.3389/fpubh.2017.00307.

18. Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, pp.283-298, DOI: https://doi.org/10.1016/S0001-2998(78)80014-2