Enhanced Knowledge in Sciences and Technology Vol. 3 No. 2 (2023) 313-323

EKST

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN: 2773-6385

The Development of Python Based Routine for Material Classification Using Laser Induced Breakdown Spectroscopy Data

Nurul Husna Mohd Adan[1], Zuhaib Haider Rizvi[1*]

[1]Department of Physics and Chemistry,
Faculty of Applied Sciences and Technology,
Universiti Tun Hussein Onn Malaysia (Pagoh Campus),
84600 Pagoh, Muar, Johor, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2023.03.02.037
Received 17 January 2023; Accepted 16 February 2023; Available online 30 November 2023

Abstract: Laser Induced Breakdown Spectroscopy (LIBS) is a highly capable tool for diverse applications. LIBS coupled with support vector machine (SVM) and random forest (RF) were developed and applied for the classification of each type of samples including gold, meat, and gemstone through Python programming language as it is open-source, and it offers a comprehensive list of libraries and packages unlike other softwares that include licensing and subscription fees. Principal component analysis (PCA) was employed on the three samples to visualize the data. The scree plot of the PCA technique has been generated, which has expressed that the first PC scores of golds, meats and gemstones were 90.1%, 96.1%, 96.4% accordingly. It has been concluded that, the Polynomial kernel from the SVM classifier has worked best for metallic samples with the classification accuracy of 89.7% while the organic samples have been well classified through the RBF kernel by SVM classifier which expressed the accuracy of 90.0%. The Polynomial kernel SVM, the RBF kernel SVM along with the RF algorithms are preferred to be utilized on the geological samples as they have shown the highest classification accuracy on the samples which were a perfect 100%.

Keywords: Laser Induced Breakdown Spectroscopy (LIBS), Principal Component Analysis (PCA), Support Vector Machine (SVM), Random Forest (RF), Classification

1. Introduction

Laser-Induced Breakdown Spectroscopy (LIBS) has developed rapidly over the last few years as a laser based analytical technique mainly due to its advantages in rapid analysis, experimental simplicity, and its ability to allow in-situ, on-line and remote elemental analysis of a sample [1]. It is a laser-based atomic emission spectroscopy (AES) technology that uses a laser-generated plasma as the vaporization, atomization, and excitation source [2]. According to this technique, it measures the emission of plasma produced as a consequence of the interaction between a highly focused laser beam and a sample. The

Corresponding author: syedzuhaib@uthm.edu.my

emitted light or as known as plasma plume includes valuable data about the elemental composition of the sample [3]. By applying LIBS technique, a lot of useful and detailed spectroscopic information from the target samples can be rapidly obtained without any prior sample treatment or preparation [4].

Machine learning approaches also have proven to be efficient in dealing with LIBS spectral data, which are often complicated and contain a great deal of noise[5]. Machine Learning (ML) is a field of artificial intelligence that uses algorithms to discover underlying correlations in data [6]. Machine learning algorithms are highly capable in retrieving valuable data from LIBS spectra, eliminating the interference, and effectively improving the classification and quantitative analysis of LIBS data with a greater precision [7]. To date, numerous methods of machine learning algorithms have been established, extending its capabilities for material classification and quantitative analysis to acquire an accurate and comprehensive analysis outcome [7]. Some of the most common ML algorithms include the Support Vector Machine (SVM), Random Forest (RF), Principal Component Analysis (PCA), Genetic Algorithm and so forth [2], [7].

Popular software packages used for such chemometric studies are MATLAB and OriginPro. These are sophisticated softwares and offer a lot more that is required for material classification using LIBS data but these softwares includes licensing and subscription fees range from USD200 to USD2,340 which is very expensive for an individual for the given application when excellent free alternatives are available. Besides, individual especially students might could not afford to pay for the fees as they do not yet have any income sources and mostly are financially supported by their parents. Furthermore, those chemometric studies which include classification analysis can be applied through a highly capable and popular programming language which is Python, that is open-source and offers a comprehensive list of libraries and packages to choose from, for chemometric and spectroscopic investigations.

Python is a programming language with a high level of object-oriented abstraction that has been widely employed for a variety of applications which works on numerous [8]. According to a recent Stack Overflow study, Python has surpassed languages such as Java, C, and C++ and has risen to the top [9]. Besides, there is a lot of research and development being conducted which covers this Python programming language [9]. It is a well-approved programming language with multiple applications ranging broadly from Artificial Intelligence (AI)-based software development to a variety of other web-based applications [10]. Furthermore, the Python framework can be used easily in building a desktop or web-based application [11]. Additionally, it is also considered as an open-source language which is free to use making it accessible to all the users [8]. The programming language and the vast majority of supporting libraries normally have flexible and open licenses [9].

The objectives of this study are to acquire LIBS data from metallic, organic, and geological samples and to perform Principal Component Analysis (PCA) on the three different kinds of samples including gold, meat, and gemstone. Other than that, this project aims to apply machine learning algorithms namely the Support Vector Machine (SVM) and the Random Forest (RF) on the gold, meat and gemstones samples for classification purposes. This project is focused on exploring the potential of two different kinds of machine learning models including the Support Vector Machine and the Random Forest in classifying a variety of materials based on their LIBS spectral data through Python programming language. The LIBS method combined with chemometric approaches is an effective technique to enhance the classification accuracy of various kinds of materials. Through this project, other researchers will be easily conducted their studies especially to those who will be working on projects related to machine learning algorithms for classification purposes to some extent as the code are able to be accessed for public.

2. Materials and Methods

Based on the Figure 1, all the datapoints which are from organic, metal, and geological samples have been obtained from the past experiments. Then, the datasets have been processed through normalization and internal standardization methods. After data processing is done, the datapoints are then have been visualized by performing PCA method on the datasets. After that, data splitting has been accomplished on the datasets by splitting the datapoints into 80% on train sets and 20% on test sets. Data splitting process is done to prevent from overfitting to occur. Overfitting by mean is when a machine learning classifier fits the training sets too well and fails in fitting the additional data. Then, all the train sets are used to train the classification models which are SVM and RF. Then, the classification models are tested by using the test datapoints. When the optimal accuracy which is greater than 98% has been achieved, the datasets is prepared to undergo the performance comparison between the classifier models and if it is not, the ratio of splitting the data must be changed. After the classification performance of both models have been compared, the results of accuracy for both machine learning classifiers are successfully obtained. All of the coding can be accessed through GitHub platform (https://gist.github.com/Annazagr).

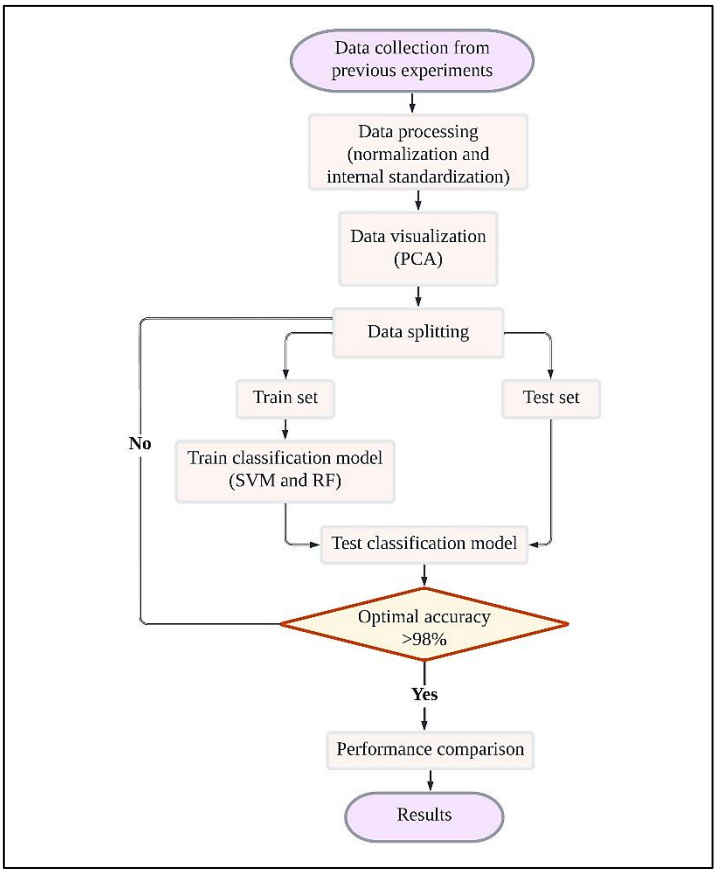


Figure 1: The main workflow diagram for data classification

2.1 Data collection

Raw LIBS spectral data from previous experiments are used in this research study to obtain the final classification results through the machine learning algorithms which has been implemented in Python. The raw LIBS spectral data have been selected from a variety of samples which are gold, gemstones and meat that represent the data from metallic, geological, and organic samples accordingly. Besides, the overall datasets have been collected from previous studies done by other researchers which have been conducted specific experiments based on material classification using LIBS technology.

2.2 Data visualization

PCA classifier model has been implemented in the coding through sklearn librabry by importing the decomposition function. Then, the mean of each value will be substracted so that all the datasets will be centered on the origin. The data have been scaled prior to feed into the PCA model through the sklearn.preprocessing.scale function. Three principal components have been selected in this analysis. The explained variance and cumulative variance have been achieved by importing the pca.explained_variance_ratio_ and the np.cumsum(np.round(explained variance)) function into the coding. After that, the PCA scores have been sorted out by applying pca.transform() function into the Python routine. Finally the data have been transformed and visualized into a scatter plot, loadings plot and scores plot through the plotly.express function.

2.3 Support Vector Machine (SVM)

The implementation of the SVM model into Python scripting requires libraries including pandas, numpy, matplotlib.pyplot, sklearn.model_selection, sklearn and mlxtend.plotting which have been imported along with the data files which are all have been saved in .csv format. the support vector classifier has been imported from sklearn.svm to define the kernel SVM classifier. The kernel SVM classifier that have been applied were Linear, Polynomial and Radial Basis Function kernels. Then, the training process has been implemented to train the kernel SVM by applying the line code of svm.svc.fit() function. After finishing the training process, the test results has been predicted by applying the svm.svc.predict( ) method. Then, the confusion matrix is being plotted by importing the accuracy_score, classification_report, and confusion_matrix function from sklearn.metrics library to evaluate the classifier performance.

2.4 Random Forest (RF)

The RF model is being implemented by importing the RF classifier module into the coding from sklearn.ensemble library package. Then, the RF classifier model has been trained to solve the classification issues. The RF classifier also uses the n_estimators function as parameter to justify the number of trees in this model. After that, the prediction process has been made on the test datasets by using the classifier.predict( ) function. Then, the RF model has been evaluated through the confusion metrics module for accuracy calculation by implementing the code for confusion_matrix and metrics.accuracy_score function from sklearn.metrics library.

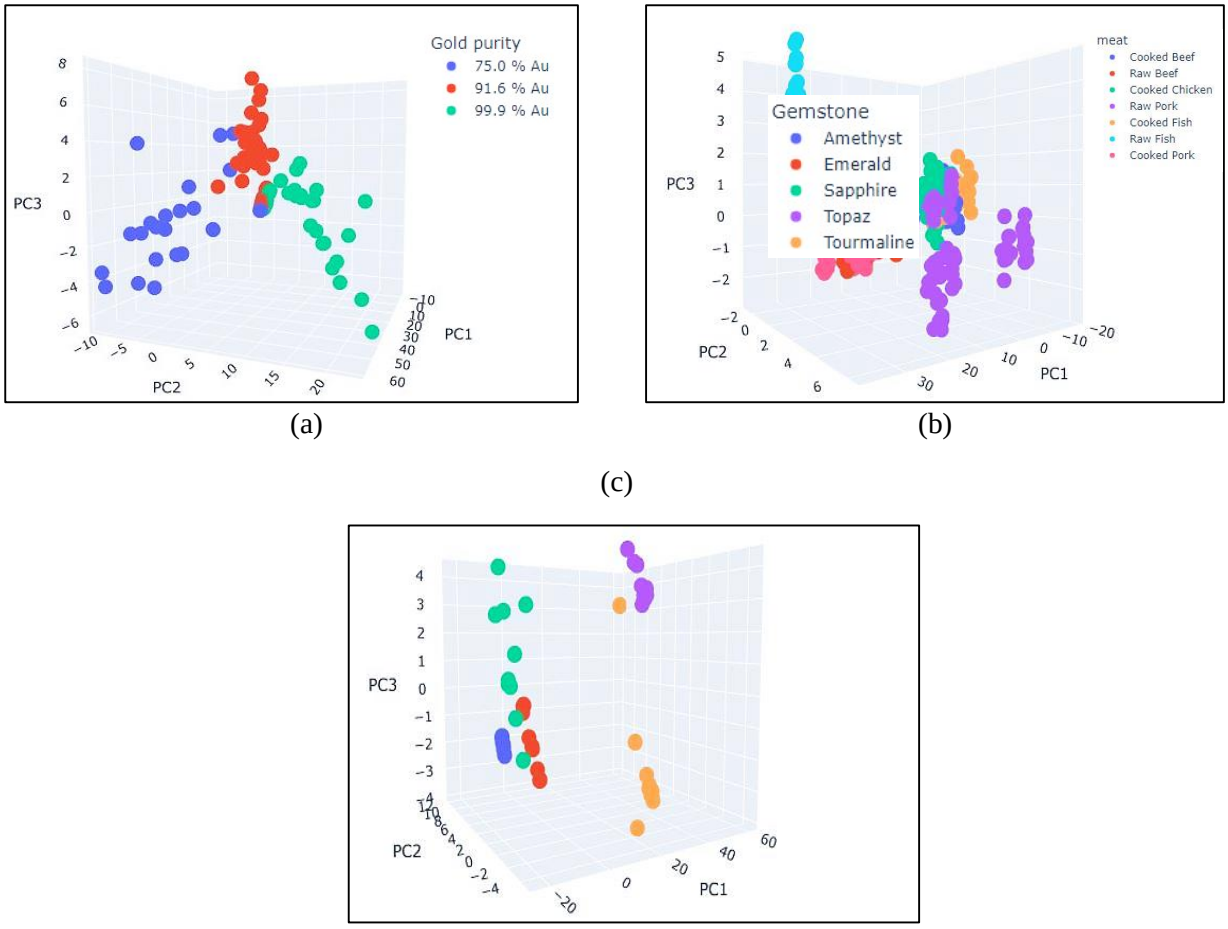
2.5 Performance comparison

The classification accuracy is defined as how near a measurement value to its actual or accepted value which is 100% for this case. Other than that, precision is described as the closeness of multiple measurement values to each other. Besides, sensitivity represents is a measure of how successfully a test can detect true positives while specificity is defined as a measure of a test's ability to detect true negatives.

3. Results and Discussion

3.1 Principal Component Analysis (PCA)

Based on Figure 2(a), almost all the variance of the gold LIBS spectral data had been compressed by PCA into three PCs. The first PC reveals the most variation which is 0.901 or 90% while the second PC captures 0.079 or 8% of variation which accounts as the second most variation reveals from the whole LIBS data sets. Cumulatively 3 PCs have 99% variance of the original dataset. It also suggests that first three PCs are adequate to interpret the overall data sets as they explain and reveal most of the variance in the gold LIBS data sets which sum up to 0.993 or 99%. This indicates how much information is being shown in respect to the raw data.



(a) (b)

(c)

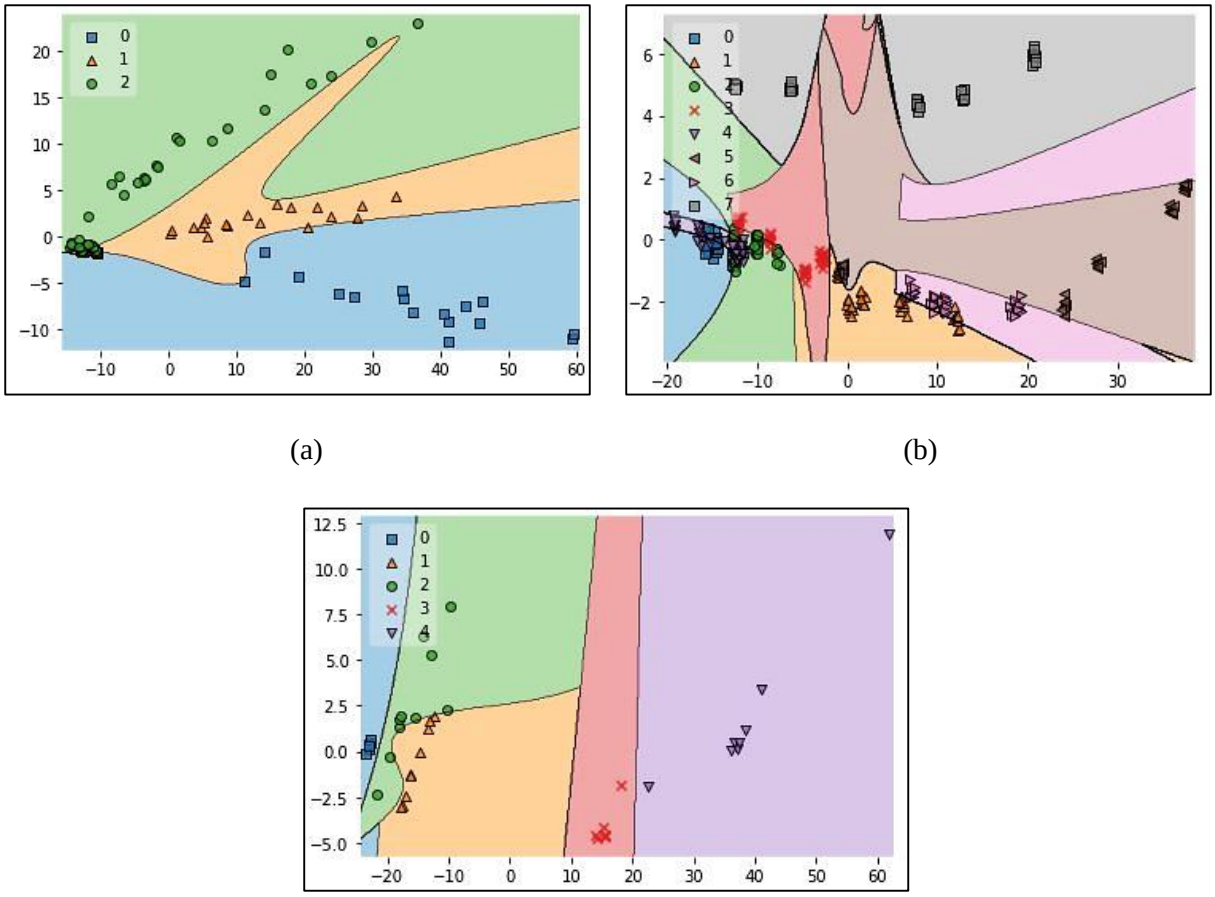
Figure 2: Scores plot of Principal Component Analysis; a) Gold data, b) Meat samples, c) Gemstones spectra

Based on Figure 2(b), it obviously shows that PCA discriminates very precisely among the different types of meat samples because most of the parts have been clustered perfectly according to the different kinds of meat which is based on its own characteristic. Therefore, it proves that PCA has the ability to clarify or summarize the raw meat LIBS spectral data so that it is simpler and easier to visualize and making the data analysis process less complicated.

Figure 2(c) represent the results achieved from the PCA of gemstone samples which are Amethyst, Emerald, Sapphire, Topaz, and Tourmaline in three-dimensional and two-dimensional view respectively. The five clusters of data are shown based on the first two principal components PC1 and PC2 that express the .96% and 2% of the cumulative variance accordingly. These five resultant clusters are well distinguished within the graph, and the point distribution is not exceptionally broad. The cluster of Amethyst, Topaz and Tourmaline are clearly separated from the mixed-up cluster of PCA scores of Sapphire and Emerald. Generally, PCA has discriminated very distinctly among these five geological samples excluding a little mixed up between the Sapphire and Emerald data.

3.2 SVM based on Polynomial kernel

Figure 3 depicts the output graph on the application of Polynomial kernel. The Polynomial kernel is frequently utilized in SVM classification technique in a condition where the data cannot be separated linearly. The Figure 3(a) shows that the gold samples are well classified through the use of SVM classifier. It can be seen clearly that the SVM model based on Polynomial kernel has performed well on gold data with an overall 89.7% accuracy.



(a) (b)

(c)

Figure 3: Results of SVM based on Polynomial kernel; a) Gold data, b) Meat spectra, c) Gemstones samples

Figure 3(b) depicts the graph of results obtained from the polynomial kernel SVM classifier implementation on the meat samples. It shows a quite clear boundary among the meat samples even there is some parts that are slightly confusing. The accuracy obtained from the employment of the polynomial kernel SVM classifier is defined to be 82.8%. The Figure 3(c) shows that the gemstones are well clustered, and it displays a clear distinction between each of the samples. The value of accuracy obtained from the polynomial kernel SVM classifier is perfect 100%.

3.3 SVM based on Linear kernel

Figure 4(a) illustrates the visual depiction of the SVM classification through the implementation of linear kernel on Gold LIBS data. It can be seen that the linear kernel SVM is performed quite efficiently as it has obviously separated the three types of golds based on their classes even though there is a little mixed up at the initial of the graph. Mainly, the linear kernel SVM is specialized for data that are linearly separable in a way that the data are able to be classified using a single line such as this gold datapoints that are happened to behave more like a linear classification issue. Therefore, the linear kernel SVM is being seen to be more efficient on this gold data. However, the accuracy achieved for this linear kernel is lower than the polynomial kernel which is 79.5%.

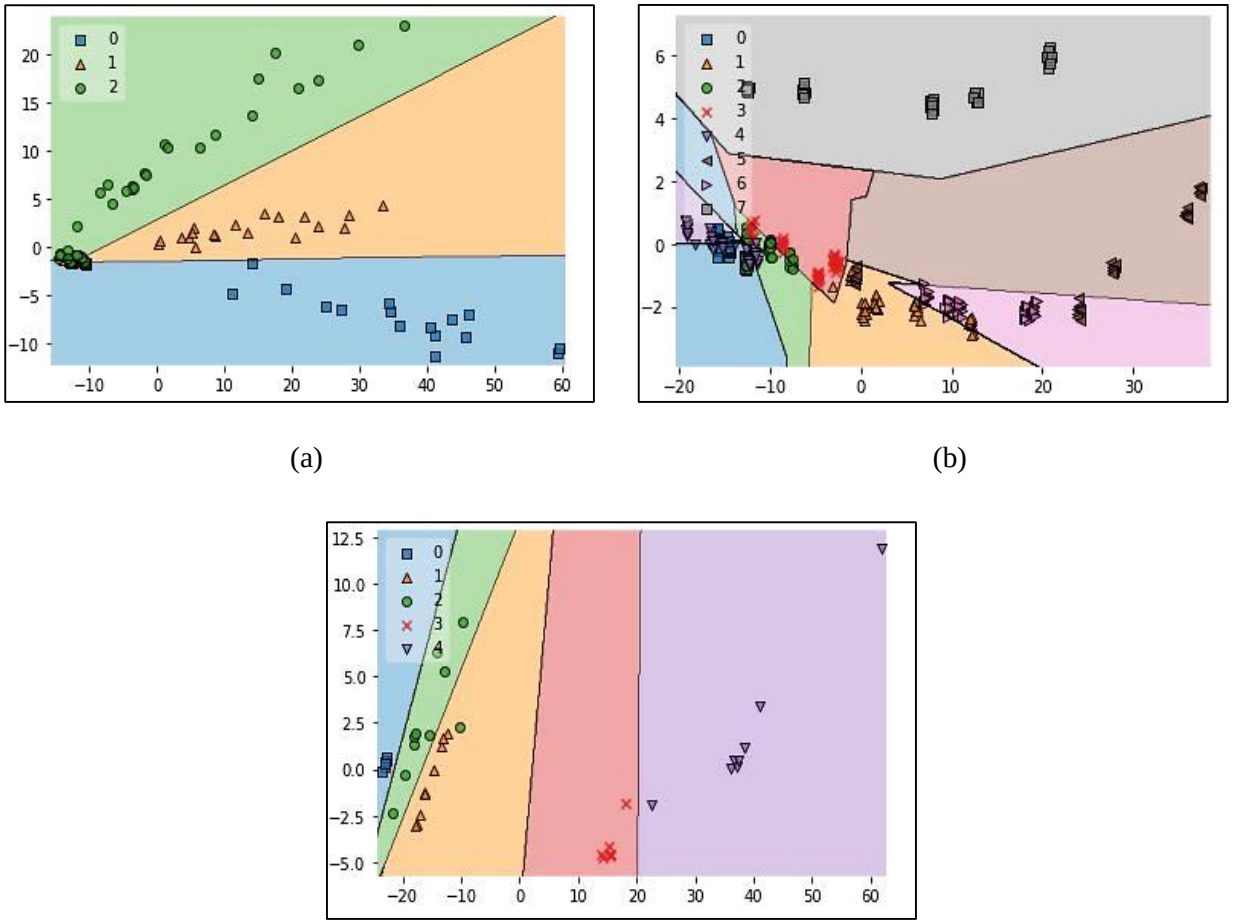
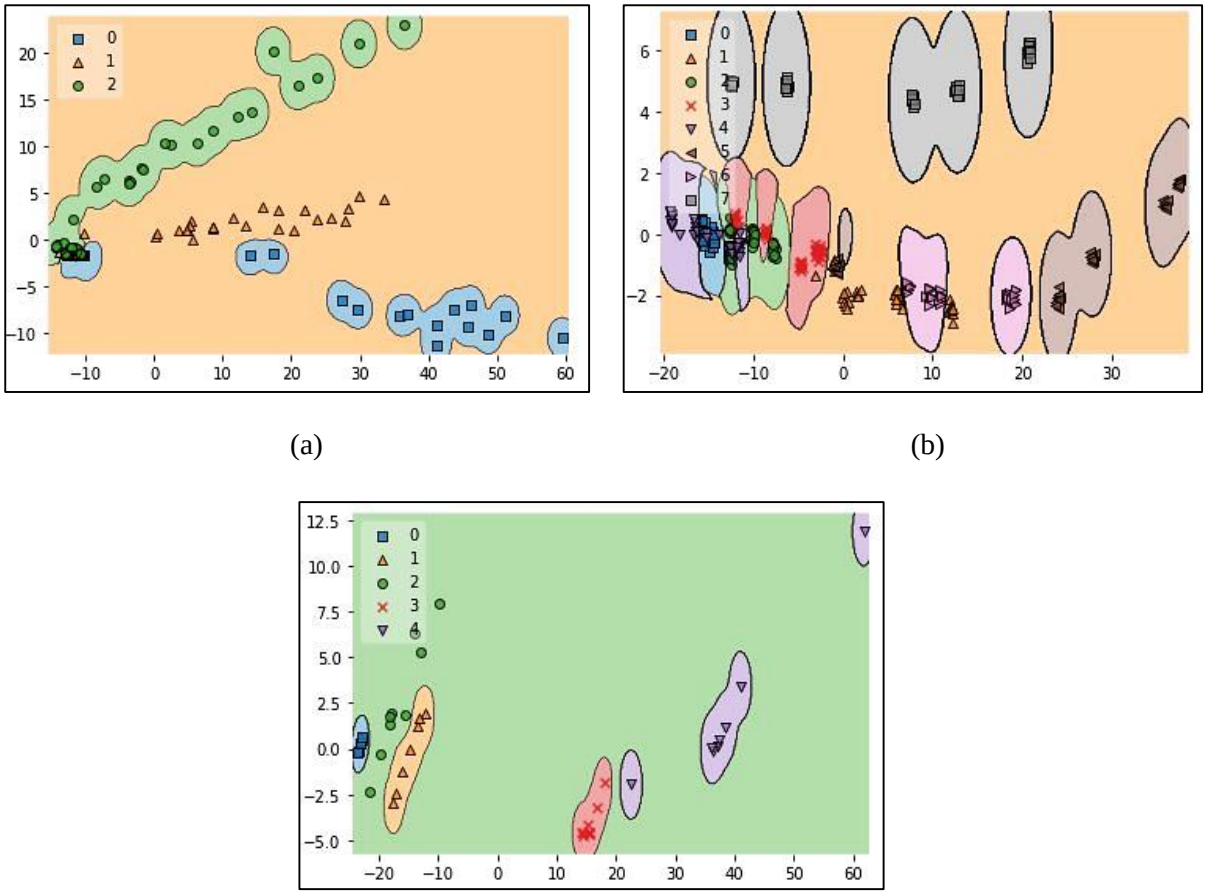


Figure 4: Results of SVM based Linear kernel classifier; a) Gold spectra, b) Meat data, c) Gemstones samples

Figure 4(b) illustrates the graph visual of the application of the linear kernel SVM on the meat samples. The accuracy achieved for this classification analysis is 76.3%. Figure 4(c) illustrates the graph generated from the implementation of the linear kernel SVM model on the gemstone samples. Distinct clusters can be clearly seen as displayed on the graph. The accuracy obtained through this linear kernel SVM method is 97.5%.

3.4 SVM based on RBF kernel

Figure 5(a) shows the results for RBF kernel on the gold samples. The RBF kernel has the ability to combine various polynomial kernels to project non-linearly distinguishable data into a higher dimensional space where it can then be separated using a hyperplane. According to the output graph of the RBF kernel, it shows that the SVM classifier has separated the datapoints based on their specific type quite efficiently. The overall accuracy obtained for the RBF kernel SVM classifier is 82.9% which indicates the second most accurate classifier after the Polynomial kernel SVM that express the overall accuracy of 89.7%. Based on Figure 5(b), it obviously depicts a clear cluster among the meat samples through the implementation of the RBF kernel SVM classifier. The overall accuracy obtained from the RBF kernel SVM classifier is 92.5% which proves that this model has performed very well in classifying the datasets. Figure 4.27 depicts the results from the utilization of the RBF kernel SVM classifier for the classification gemstone samples. It shows that the datapoints have been completely separated through clear boundaries between the different kinds of samples. It proves that the RBF kernel SVM model is performing very well with the accuracy of 100%.



(a) (b)

(c)

Figure 5: Support Vector Machine based RBF kernel classifier; a) Gold spectra, b) Meat samples, c) Gemstones data

3.5 Random Forest (RF)

The RF model has correctly predicted 15 data for Au 75.0% while 2 of the data have been mis-categorized in the Au 91.6% class. Besides, a total of 10 of Au 91.6% samples have been precisely predicted out of 12 of the Au 91.6% samples. On the other hand, the RF model has accurately predicted all of Au 99.9% samples without any misidentification. The application of RF algorithm on gold LIBS spectra has demonstrated the accuracy of 89.2%.

A total of 72 samples data are correctly predicted out of the total 80 meat samples by the RF algorithm. This implies that, the RF classifier has performed very well in predicting the 8 different types of meat spectral data with an overall accuracy of 90.0% and mapping the data to which they belong in the confusion matrix.

The gemstone samples data including Amethyst, Emerald, Sapphire and Topaz have been predicted accurately by the RF model without any misclassification. A total of 11 instances have been mapped precisely according to their belonging class as expressed in the confusion matrix. Thus, it proves that the RF classifier has worked very efficiently on the gemstone samples with accuracy of perfect 100%.

3.6 Performance comparison

Table 1 depicts the classification accuracy of two kinds of ML including the RF and the SVM classifier with three different kernels namely Polynomial, Linear and RBF on three types of samples which are golds, meats and gemstones. It is noticed that the SVM model by Polynomial kernel method has the highest classification accuracy for gold samples compared to other ML techniques. Diaz Romero et al, have adopted ML algorithm such as SVM to analyze metal LIBS spectral data for the classification of aluminium metal scrap samples and have proven that SVM coupled with LIBS spectral data have expressed higher classification accuracy compared to RF approach. On the other hand, meat samples have been well classified through the implementation of the SVM model by RBF kernel with the classification accuracy of 92.5%. This indicates that, the RBF kernel SVM model works efficiently for meat samples contrasted to the RF and other SVM kernel classifiers. In the work of Zhan et al, they have reported that the RBF kernel SVM model has functioned well in classifying the fish meat samples with the classification accuracy of 91.5%. Furthermore, three classifiers namely the RF, the Polynomial kernel SVM and the Linear kernel SVM have shown a great performance in the classification of gemstones LIBS spectra with the highest accuracy of 100%. This indicates that the gemstones samples are easiest to be classified through the application of the RF, the RBF kernel SVM and the Polynomial kernel SVM models. High accuracy of ML models obtained for gemstones data is because of the size of the datasets which is below 50 data points as compared to meat and gold samples that is above 50 data points. Figure 6 illustrates the performance comparison of ML approaches on the three different types of data samples including gold, meat, and gemstones with the aid of bar charts.

Table 1: Classification accuracy of Machine Learning (ML) models

	Classification Accuracy (%)			
Types of Samples	Polynomial (SVM)	Linear (SVM)	RBF (SVM)	Random Forest
Gold	89.7	79.5	82.9	89.2
Meat	82.8	76.3	92.5	90.0
Gemstones	100.0	97.5	100.0	100.0

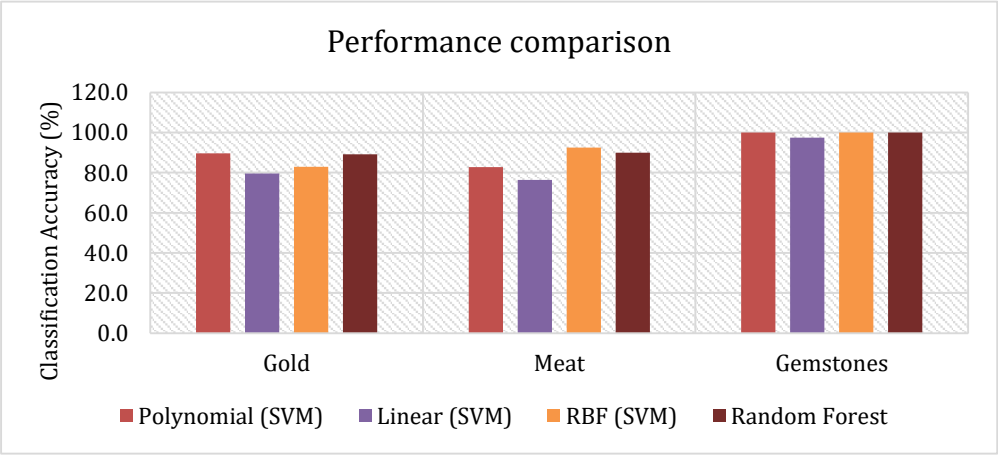


Figure 6: Performance comparison among SVM and RF classifiers

4. Conclusion

In summary, the objectives defined in this study have been achieved successfully. The PCA approaches for data visualization has been developed through Python programming language and has

been employed for all types of samples. The datapoints for all samples have been visualized excellently through the scores plot of PCA in 2D and 3D views. The scree plot of PCA method has been generated, which provides the first PC scores of 90.1%, 96.1%, 96.4% for gold, meat, and gemstones samples respectively. Furthermore, the RF and the SVM models have been successfully executed on the samples. It has proven that, the Polynomial kernel from SVM model has worked best for metallic samples while the organic samples have been well classified through the RBF kernel by SVM classifier. On the other hand, the RF algorithm are preferred to be utilized on geological samples as it has shown the highest classification accuracy on the gemstone samples. In General, the Python code has generated similar results as in the origin or matlab with the advantages of no subscription fees needed. Therefore, this study will be beneficial and cost effective for others that would have wanted to apply the ML algorithms within their projects.

Acknowledgement

This study was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through vote H848. The authors would also like to thank the Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia for its support.

References

[1] E. Bellou, N. Gyftokostas, D. Stefas, O. Gazeli, and S. Couris, "Laser-induced breakdown spectroscopy assisted by machine learning for olive oils classification: The effect of the experimental parameters," Spectrochim Acta Part B At Spectrosc, vol. 163, Jan. 2020.

[2] N. Gyftokostas, E. Nanou, D. Stefas, V. Kokkinos, C. Bouras, and S. Couris, "Classification of Greek olive oils from different regions by machine learning-aided laser-induced breakdown spectroscopy and absorption spectroscopy," Molecules, vol. 26, no. 5, Mar. 2021.

[3] R. Gaudiuso et al., "Laser-induced breakdown spectroscopy for human and animal health: A review," Spectrochimica Acta - Part B Atomic Spectroscopy, vol. 152. Elsevier B.V., pp. 123–148, Feb. 01, 2019.

[4] D. Stefas, N. Gyftokostas, and S. Couris, "Laser induced breakdown spectroscopy for elemental analysis and discrimination of honey samples," Spectrochim Acta Part B At Spectrosc, vol. 172, Oct. 2020.

[5] G. Marinova and M. Todorova, "Classification Tasks Solving with Machine Learning Methods," in 2020 29th International Scientific Conference Electronics, ET 2020 - Proceedings, Sep. 2020.

[6] A. M. Sequeira, D. Lousa, and M. Rocha, "ProPythia: A Python Automated Platform for the Classification of Proteins Using Machine Learning," in Advances in Intelligent Systems and Computing, 2021, vol. 1240 AISC, pp. 32–41.

[7] T. Chen, T. Zhang, and H. Li, "Applications of laser-induced breakdown spectroscopy (LIBS) combined with machine learning in geochemical and environmental resources exploration," TrAC - Trends in Analytical Chemistry, vol. 133. Elsevier B.V., Dec. 01, 2020.

[8] D. Lafuente et al., "A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling," J Chem Educ, vol. 98, no. 9, pp. 2892–2898, Sep. 2021.

[9] A. L. S. Saabith, M. Fareez, and T. Vinothraj, "Related papers Python Current Trend Applications-an Overview Popular Web Developement Frameworks In Python," International Journal of Advance Engineering and Research Development, vol. 6, no. 10, 2019.

[10] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthcare Analytics, vol. 2, p. 100016, Nov. 2022.

[11] P. Mathur, Machine learning applications using python: Cases studies from healthcare, retail, and finance. Apress Media LLC, 2018.