Machine Learning Approach for Material Analytics and Classification – Insights Based on a Criminal Forensic Investigation Data

K. Venkatesh Raja[1*] [ID], B. Ayshwarya[2] [ID], D. Mohana Geetha[3] [ID], V. Nagaraj[4] [ID], R. Ramkumar[5] [ID]

[1] Department of Mechanical Engineering, Sona College of Technology, Salem 636005, Tamil Nadu, India
[2] Department of Computer Science, Kristu Jayanti College, Bengaluru 560077, Karnataka, India
[3] Department of Electronics & Communication Engineering, Sri Krishna College of Engineering & Technology, Coimbatore 641008, Tamilnadu, India
[4] Department of Electronics & Communication Engineering, Annapoorana Engineering College, Salem 636308, Tamil Nadu, India
[5] Department of Electronics & Communication Engineering, Bannari Amman Institute of Technology, Erode 638401, Tamil Nadu, India

Corresponding Author Email: kvenkateshraja@hotmail.com

ABSTRACT

Glass is a non-crystalline chalcogenide amorphous solid that is often transparent and has widespread practical, technological, and decorative usage in, for example, windowpanes, tableware, and optoelectronics. Each type of glass has different material compositions to better suit the required application. Composition of glass has various material compositions like Si, Na, Mg, Al, Ca, Ba and so on, based of which the type of glass is classified. This research work primarily focuses on assessing the capability of Machine Learning models for predicting the type of glass left in a crime scene which can further be utilized for higher levels of criminological investigations. The proposed research process incorporates collection of data set from records of forensic investigation. Further, the data is processed for any errors and processed with the aid of popular machine learning algorithms viz. Regression, decision trees, k-means clustering and random forest classifier. The proposed data set has seven different types of glass attributes with 224 sample instances are used in this study for classification. From the results it is evident that, random forest algorithm performs well with higher magnitudes of accuracy.

1. INTRODUCTION

The fragments of glass material left in a crime spot always remains as a wonderful clue for the criminal investigator to move further in detecting the root cause of the issue. All the fragments are usually collected and the same may give some information regarding the sequence of incidents that happened during the time of crime. Even small fragments of glass analysis may reveal some big clues for further investigations. Aldayel [1] employed k-nearest algorithm (kNN) for identifying the type of glass present in the criminal scene. The algorithm predicted and classified 172 data correctly out of total 214 instances present in the data base with an accuracy of about 80%. A new classification system was proposed to identify ceramic, and ceramic based restorative materials based on the phase or phases available in their chemical composition [2]. Baird et al. [3] developed a predictive method based on machine learning approach is devised to identify the properties of inorganic crystalline and amorphous materials. The characteristics of a group of 51 organic molecules are evaluated and classified based on the amorphous formulation of materials [3]. Later, Tiwari et al. [4] investigated the fragments of glass pieces left in the crime based on fractography approach. Further, an interesting study was reported to identify the glass type based on the fracture pattern to add value to forensic community [5]. Recently Sun et al. [6]

developed a machine learning approach to predict the glass forming ability with the aid of support vector classification technique. The results add significant boost to use Machine Learning (ML) approach for such kind of prediction and identification problems. Furthermore, the efficiency of deploying ML tools in prediction and understanding the characteristics of amorphous bulk metal glasses were presented in detail by Xiong et al. [7]. Few other applications of Machine Learning algorithms in material testing segment which includes failure load estimation, parameter estimation in deep hole drilling process, computation of surface roughness and certain optimization applications were reported [8-10]. Recently, Yuliawan et al. [11] deployed machine learning models to predict and classify the socio-economic vulnerability factors. Kazemi and Niaki [12] applied classification techniques for monitoring image-based processing systems. Moreover, certain studies on forensic science is also reported by Margagliotti and Bollé [13].
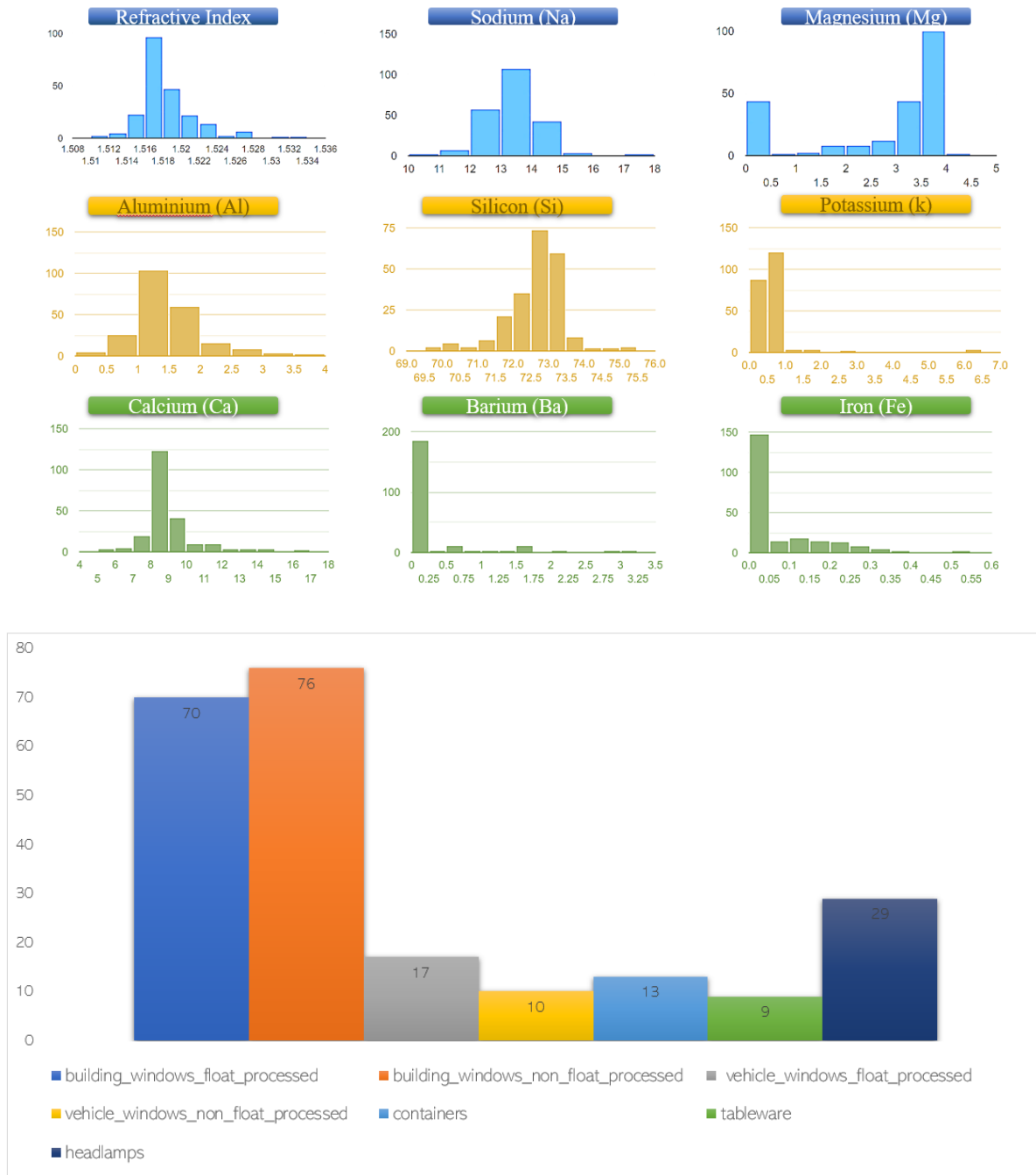
However, only negligible quantum of works was reported to predict the type of glass fragments left in a crime scene based the material composition and characteristics of the glass material. Based on the above facts, this research work is devoted to adding value to forensic investigation of glass material based on the refractive index and the composition of glass material. Popular ML methods are to be tested with the data set and the results would attract the interest of forensic

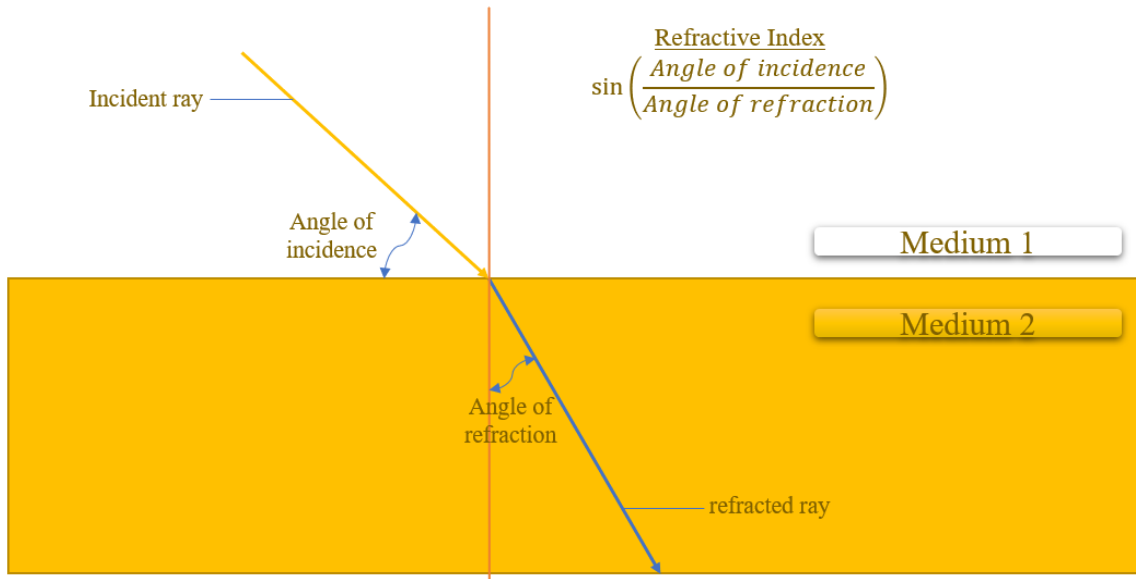investigators to apply these techniques for real time case studies.

2. DATA SET – METRICS

For investigation and analysis, the data set is taken from UCI Machine Learning Repository having 214 instances developed by Home Office Forensic Science Service, Berkshire [14]. The parameters and range of values presented in the data set is clearly exhibited in the Table 1 for reference and further assessment. The box and whisker plot clearly represents the distribution pattern of all parameters with outliers. Also, the table exhibits the statistical factors (mean, Standard deviation (SD), the maximum and min values) for

reference. For more clarity, the range of distribution factors are portrayed in Figure 1 for all the parameters used in this study. In the data set, nine input parameters were utilized in deciding the output parameter. Here, the first factor refractive index is a terminology that is represented by a dimensionless number which visualises the bending ability of light rays while passing through different mediums. The refractive index is computed as the ratio between sine angle of incidence and sine angle of refraction. Moreover, the same parameter can be related to velocity of light with reference to its velocity in a particular medium (Figure 2). Other input parameters ranging from 2 to 9 represents its weight percentage in corresponding oxide formulation.

Figure 1. Distribution ranges of the input and output parameters

$$\text{Refractive Index} \quad \sin\left(\frac{\text{Angle of incidence}}{\text{Angle of refraction}}\right)$$

Figure 2. Refractive index computation

Table 1. Glass data parameter description

ID	Parameter	Statistical factors				Distribution pattern
		Min	Max	Mean	SD	
1	Refractive Index (RI)	1.51	1.53	1.52	0.0030	
2	Sodium (Na)	10.73	17.38	13.41	0.8166	
3	Magnesium (Mg)	0	4.49	2.68	1.4424	
4	Aluminium (Al)	0.29	3.5	1.44	0.4993	
5	Silicon (Si)	69.81	75.41	72.65	0.7745	
6	Potassium (K)	0	6.21	0.48	0.6522	
7	Calcium (Ca)	5.43	16.19	8.96	1.4232	
8	Barium (Ba)	0	3.15	0.18	0.4972	
9	Iron (Fe)	0	0.51	0.05	0.0974	

The type of glass material is classified by numerals from 1 to 7 (1-building windows float processed, 2-building windows non float processed, 3-vehicle windows float processed, 4-vehicle windows non float processed, 5-containers, 6-

tableware and 7-headlamps). Since the database is not having any instance of non-float processed vehicle windows, 10 new instances of the same category are being added within the distribution limits of material compositions with a view to enhance the prediction quality in a realistic fashion. Hence, the new data base instance is increased to 224. The distribution of glass variations used in this research work is depicted in Figure 3 for reference.
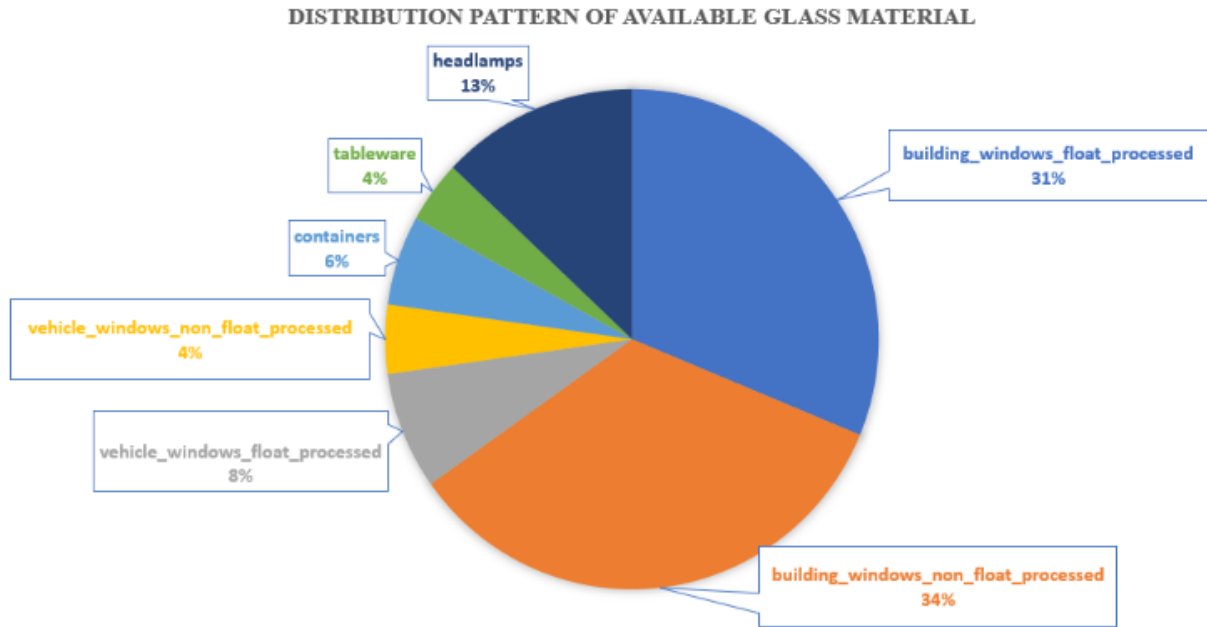

Figure 3. Class distribution of the modified glass data set with 224 instances

3. ML SCHEMA AND METHODOLOGY

Seven popular Machine Learning algorithms (Linear regression, polynomial regression, k-means clustering, kNN technique, decision tree system, support vector machine and random forest algorithm) are selected for analysing the data pattern based on the proven robustness and versatile application domain. To simplify the coding schema Scikit-Learn [15] libraries are installed in Anaconda python programming module and necessary libraries are imported at the time of coding. Scikit-Learn [15] libraries have the capability of hyper parameter tuning and ease of coding methodology. The performance of the algorithms is measured by its prediction capability and evaluated in terms of correlation parameter metrics. The data set is bifurcated in two categories: one for training and hyper parameter tuning and the next one for testing in the ratio of 75:25. The schema of the proposed methodology is clearly illustrated in Figure 4.

4. RESULT ANALYSIS

To ensure the accuracy and precision of results presented in this section, each algorithm is executed for 10 runs and the mean value is reported. Also, the variation of hyper-parameters in glass material prediction is discussed. The results are presented in the form of heat-map variants and precision-recall curves for better clarity along performance metrics. As depicted in the abstract section seven different Machine Learning algorithms are used in this study. The data set is being coded and executed with the proposed algorithms using Python programming module. The hyper-parameters of the algorithms are tuned in such way to deliver more optimized results.
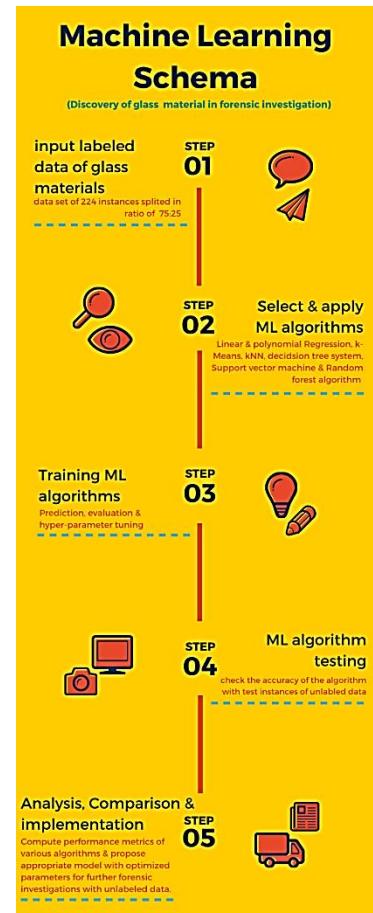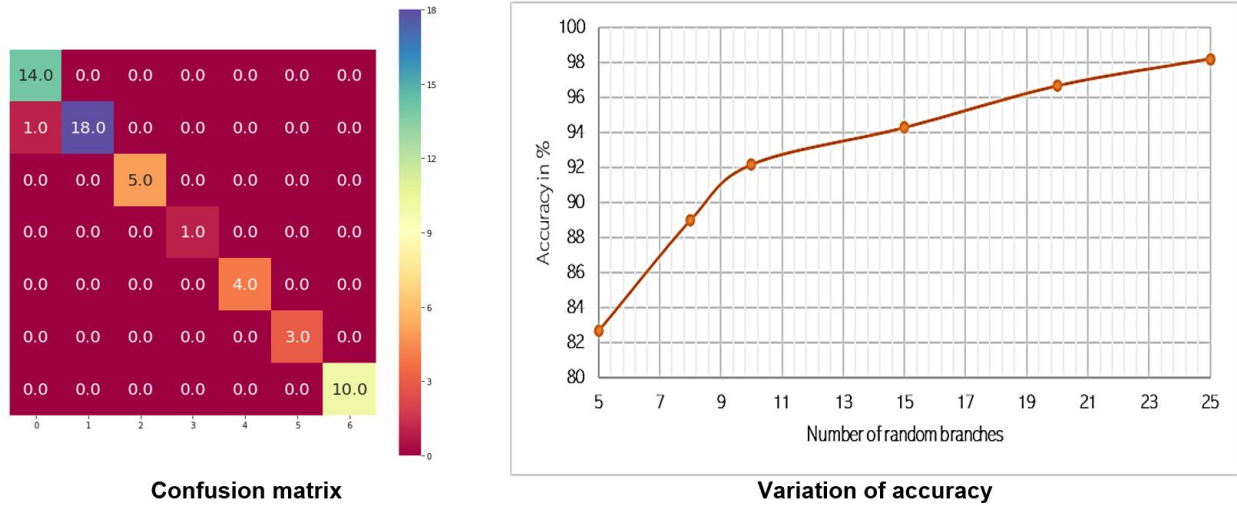

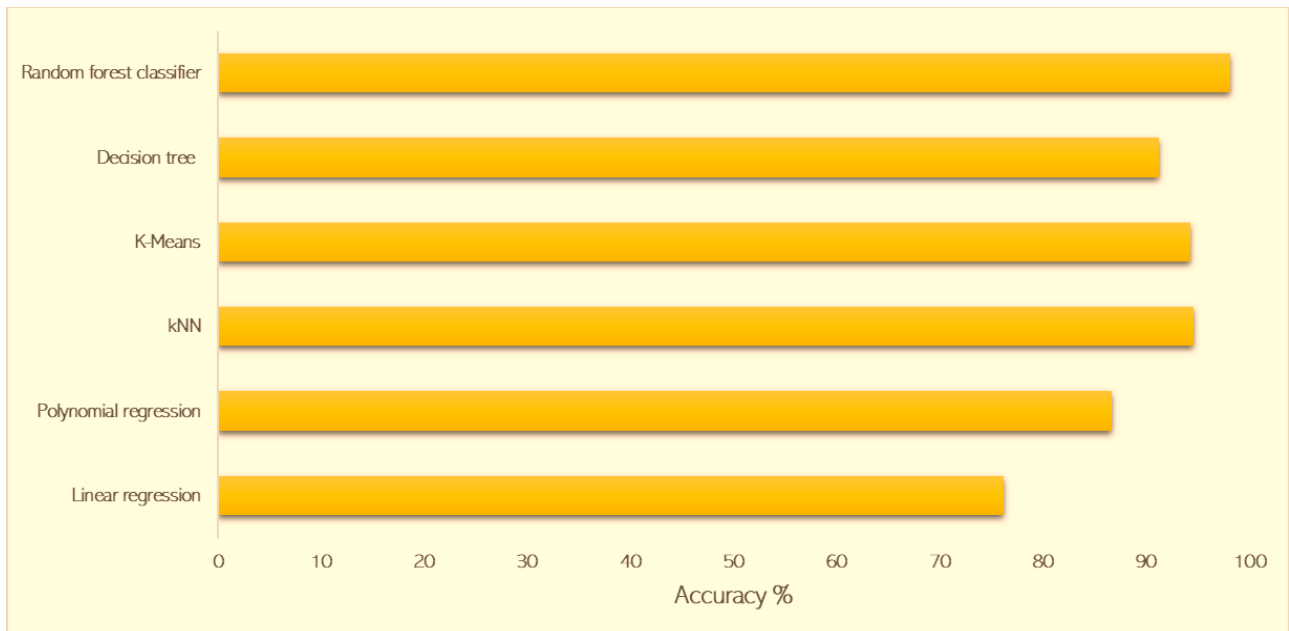**Figure 4.** Schema of machine learning methodology

Figure 5 highlights the confusion matrix and predicted results based on random forest classifier (RFC). Confusion matrix is a term used in all ML techniques where-in the prediction capacity can be assessed in an easier and understandable fashion. From the figure it is evident that RFC technique is robust in predicting the results while compared with all other algorithms. As the number of branches in the RFC is increased (Figure 5), naturally the accuracy prediction is also enhanced. Also, the computational cost associated with the algorithm is extremely lower while compared with all other techniques. The x-values given in the confusion matrix (0 – 6) represents the type of glass material described in Figure 1.

Figure 6 reports the overall comparison of all selected algorithms in terms of its prediction capability. Linear regression model is not suitable for data sets with multiple attributes. However, polynomial model has better regression characteristics due to higher degree of polynomial. K-means clustering and kNN algorithm is also suitable and robust for low volume data base.

Confusion matrix

Variation of accuracy

Figure 5. Performance metrics of RFC with 98.21 % accuracy

Figure 6. Overall comparison of ML algorithms

5. CONCLUSION

This research work explains the application of various ML algorithms that can be applied to detect the type of glass material left in a crime scene. Identification of glass is carried out based on material composition and refractive index. This research work will cater to the needs of forensic investigators to move towards the criminal investigation at an enhanced dimension by applying Machine Learning techniques. Also, the same concept can be utilized to categorize or classify materials based on various characterization and analysis. Limitations of this study includes the applications to large data sets, accuracy challenges incurred due to errors in the raw data set collection. As the world is moving towards a digital dimension, incorporation of such modern tools might enhance the clarity of the results derived by a forensic investigator. Future works can be extended towards image processing and dynamic data detection.

CONFLICT OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest in publication of this manuscript.

REFERENCES

[1] Aldayel, M.S. (2012). K-Nearest neighbor classification for glass identification problem. In 2012 International Conference on Computer Systems and Industrial Informatics, IEEE, Sharjah, United Arab Emirates. https://doi.org/10.1109/ICCSII.2012.6454522

[2] Gracis, S., Thompson, V.P., Ferencz, J.L., Silva, N.R.F.A., Bonfante, E.A. (2016). A new classification system for all-ceramic and ceramic-like restorative materials. International Journal of Prosthodontics, 28(3): 227-235. https://doi.org/10.11607/ijp.4244

[3] Baird, J.A., Van Eerdenbrugh, B., Taylor, L.S. (2010). A classification system to assess the crystallization tendency of organic molecules from undercooled melts. Journal of Pharmaceutical Sciences, 99(9): 3787-3806. https://doi.org/10.1002/jps.22197

[4] Tiwari, N., Harshey, A., Das, T., Abhyankar, S., Yadav, V.K., Nigam, K., Anand, V.R., Srivastava, A. (2019). Evidential significance of multiple fracture patterns onthe glass in forensic ballistics. Egyptian Journal of Forensic Sciences. https://doi.org/10.1186/s41935-019-0128-4

[5] Harshey, A., Srivastava, A., Yadav, V.K., Nigam, K., Kumar, A., Das, T. (2017). Analysis of glass fracture pattern made by .177" (4.5 mm) Caliber air rifle. Egyptian Journal of Forensic Sciences. https://doi.org/10.1186/s41935-017-0019-5

[6] Sun, Y.T., Bai, H.Y., Li, M.Z., Wang, W.H. (2017). Machine learning approach for prediction and understanding of glass-forming ability. The Journal of Physical Chemistry Letters. https://doi.org/10.1021/acs.jpclett.7b01046

[7] Xiong, J., Shi, S.Q., Zhang, T.Y. (2020). A machine-learning approach to predicting and understanding the properties of amorphous metallic alloys. Materials & Design. https://doi.org/10.1016/j.matdes.2019.108378

[8] Ozkan, M.T. (2015). Surface roughness during the turning process of a 50CrV4 (SAE 6150) steel and ANN based modeling. Materials Testing, 57(10): 889-896. https://doi.org/10.3139/120.110793

[9] Ugur, E., Kulekci, M.K., Ozgun, S., Kazancoglu, Y. (2013). Predictive modelling of ball burnishing process using regression analysis and neural network. Materials Testing, 55(3): 187-192. https://doi.org/10.3139/120.110423

[10] Al-Wedyan, H.M., Hayajneh, M.T. (2017). Prediction and controlling of roundness during the BTA deep hole drilling process: Experimental investigations and fuzzy modeling. Materials Testing, 59(3): 284-289. https://doi.org/10.3139/120.110999

[11] Yuliawan, D., Hakim, D.B., Juanda, B., Fauzi, A. (2022). Classification and prediction of rural socio-economic vulnerability (IRSV) integrated with social-ecological system (SES). Decision Science Letters, 11(3): 223-234. http://dx.doi.org/10.5267/j.dsl.2022.4.001

[12] Kazemi, S., Niaki, S.T.A. (2021). Monitoring image-based processes using a PCA-based control chart and a classification technique. Decision Science Letters, 10(1): 39-51. http://dx.doi.org/10.5267/j.dsl.2020.10.005

[13] Margagliotti, G., Bollé, T. (2019). Machine learning & forensic science. Forensic Science International. https://doi.org/10.1016/j.forsciint.2019.02.045

[14] UCI Machine Learning Repository. (1987). Glass Identification Dataset. http://archive.ics.uci.edu/ml/datasets/Glass+Identification

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.