

# Regression-Based Mathematical Framework for Raman Spectra of Oxide Glasses

<sup>1</sup>M A Haleem Rizwan, <sup>2</sup>P. Suresh

<sup>1</sup>Research scholar, Department of Mathematics, Dravidian University, Kuppam, A.P, India

<sup>2</sup>Asst.Professor, Department of Mathematics, CBIT, Hyderabad, Telangana State, India

Email – <sup>1</sup>haleemrizwan@mjcollege.ac.in , <sup>2</sup>suresh.pallerla000@gmail.com

**Abstract:** *The present work first time investigates the application of regression techniques, specifically Support Vector Machine (SVM) and Random Forest algorithms, in predicting the Raman spectra of borosilicate glasses with the composition  $50\text{SiO}_2-(15-x)\text{B}_2\text{O}_3-20\text{Na}_2\text{O}-10\text{ZnO}-5\text{ZrO}_2-x\text{La}_2\text{O}_3$ . Raman spectra analysis showed shifts and intensity changes particularly the conversion of  $\text{BO}_3$  to  $\text{BO}_4$  units and shifts in peak intensities and positions, indicating modifications in the glass network due to varying  $\text{La}_2\text{O}_3$  content. Both SVM and Random Forest algorithms exhibited high accuracy in predicting Raman spectra, with precision, recall, and F1 scores ranging between 0.958-0.969, 0.967-0.975, and 0.988-0.997, respectively. The Random Forest model, in particular, provided highly accurate predictions ( $R^2 = 0.978$ ), outperforming the SVM model ( $R^2 = 0.925$ ). These findings highlight the potential of regression techniques in advancing glass science for designing glasses tailored to specific applications.*

**Key Words:** *Raman Spectra, SVM Regression, Random Forest regression.*

## 1. INTRODUCTION:

The significance of glass properties influenced by network formers and modifiers, particularly in borosilicate glasses. Network modifiers, existing as single ions within the cross-linked network, alter properties such as melting point, viscosity, and thermal/electrical characteristics [1,2]. This is crucial for the chemical durability of glasses, especially in the context of immobilizing high-level radioactive waste (HLW) materials. HLW, generated from reprocessed nuclear fuel, contains various actinides and fission products requiring long-term storage in inert host materials [3,4]. Borosilicate glasses are favored for this purpose due to their mechanical and chemical resilience. Research worldwide focuses on modeling the effects of incorporating radioactive constituents into glass structures, determining solubility limits, and studying the durability and stability of potential host materials [5].

In recent years, mathematical regression techniques have seen increasing application in glass science [6-11]. These methods have been utilized to predict various properties of oxide glasses and to identify material compositions through spectroscopic analysis. Typically, Raman spectra contain valuable chemical information alongside baselines and random noise. However, these latter components can hinder the accuracy of qualitative substance analysis. Fortunately, Regression techniques algorithms offer an effective means of mitigating this issue. The prediction of Raman spectra for oxide glasses represents a novel frontier in the field of glass science using machine learning (ML). This study marks the first instance of such predictions being reported for Raman spectra of glasses. In the current study, borosilicate glasses with the composition  $50\text{SiO}_2-(15-x)\text{B}_2\text{O}_3-20\text{Na}_2\text{O}-10\text{ZnO}-5\text{ZrO}_2-x\text{La}_2\text{O}_3$  were investigated. Raman spectra of present glasses is examined and predicted using a range of regression techniques

## 2. DATA SET:

The prediction of Raman spectra for the current glasses relies on a dataset comprising Raman spectra from 100 borosilicate samples. Each sample contains approximately 1000 data points, resulting in a total of 100,000 data points considered for the regression models in this study. Figure 1 illustrates the Raman spectra of various glass samples from the dataset under investigation.

## 3. REGRESSION TECHNIQUES:

### (a) Decision Tree Regression

Decision Tree Regression is a non-parametric supervised learning method used for both classification and regression tasks. Here, I'll explain Decision Tree Regression specifically for predicting continuous values (regression).

### Mathematical Formulation

Decision Tree Regression works by recursively partitioning the feature space into regions and predicting the average (or weighted average) of the target values of the training samples in each region.

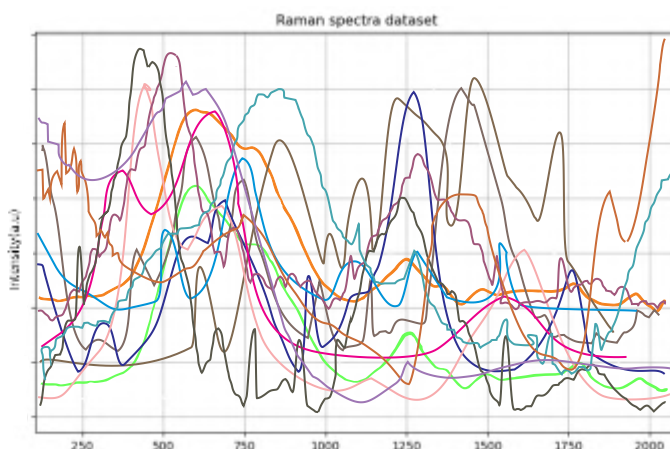


Figure 1 Raman spectra of various glass samples

### Basic Concept

Given training data  $\{(x_i, y_i)\}_{i=1}^N$  Where  $x_i \in R^p$  are the feature vectors and Where  $y_i \in R$  are the corresponding target values, Decision Tree Regression aims to partition the feature space into disjoint regions  $R_1, R_2, \dots, R_M$  each region  $R_m$  is associated with predicted value  $\hat{y}_m$

Mathematical Representation:

The prediction for a new input  $x$  using decision tree  $T$  can be represented mathematically as:

$$\hat{y} = T(x)$$

Where  $T(x)$  denotes the prediction for input  $x$  using decision tree  $T$

Splitting Criterion:

Typically, the splitting criterion in decision trees for regression is based on minimizing the variance of the target values in each region. For instance, one common criterion is to minimize the mean squared error (MSE):

Squared error (MSE)

$$MSE = \frac{1}{N_m} \sum_{i \in R_m} (y_i - \bar{y}_m)^2$$

$N_m$  is the number of samples in region  $R_m$

$y_i$  are the observed values of the target variable

$\bar{y}_m$  is the mean of the target values in region  $R_m$

### (b) Random forest Regression

A collection of decision trees where each tree gives a prediction and the final prediction is the average (regression) or majority (classification) of individual tree predictions. Random Forest Regression is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees for regression tasks. Here's an overview of Random Forest Regression mathematically:

### Mathematical Formulation

Random Forest Regression builds upon Decision Tree Regression by creating an ensemble of decision trees and averaging their predictions. Each tree is trained independently on a random subset of the data and a random subset of the features.

Ensemble learning

Given training data  $\{(x_i, y_i)\}_{i=1}^N$  Random Forest Regression creates  $B$  decision trees

$T_1, T_2, \dots, T_B$

### Training

For each tree  $T_b$  (where  $b = 1, 2, \dots, B$ )

1. Randomly select a subset of the training data (bootstrapping).
2. Randomly select a subset of features to use for splitting at each node.

### Prediction

To predict the target value  $y^{\wedge}$  for a new input vector  $x$ :

Aggregate the predictions of all trees  $T_b$ :

$$y^{\wedge} = \sum_{b=1}^B T_b(x)$$

Mathematical Representation:

The prediction for Random Forest Regression can be mathematically represented as the average of predictions from individual decision trees

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

Where

$\hat{y}$  is the predicted value

$T_b(x)$  Denoted the prediction of the  $b$ -th decision tree for input vector  $x$

$B$  is the number of trees in the forest

### Support Vector Regression (SVR):

Support Vector Regression (SVR) is a regression technique that uses Support Vector Machines (SVMs) to find the best fitting line (or hyperplane in higher dimensions) in a high-dimensional feature space. SVR is particularly useful when dealing with non-linear relationships between variables. Here's an elaboration of Support Vector Regression mathematically:

#### Mathematical Formulation:

SVR builds upon the principles of SVM for classification and extends it to regression problems. The goal of SVR is to find a function  $f(x)$  that predicts a continuous target variable  $y$  based on input features  $x$ .

Basic Concept:

Given training data  $\{(x_i, y_i)\}_{i=1}^N$

Where  $x_i \in R^p$  and  $y_i \in R$  SVR seeks to find a function  $f(x)$

Such that:

$$y_i = f(x_i) + \epsilon_i$$

Where  $\epsilon_i$  are the errors or residuals, subject to certain constraints

Formulation:

SVR introduces a margin of tolerance  $\epsilon$  around a fitting hyperplane in the feature space. The basic SVR formulation aims to minimize the complexity of the model (i.e., the norm of the weights) subject to the error tolerance  $\epsilon$ :

$$\min_{w, \zeta, \zeta^*} \frac{1}{2} w^T w + c \sum_{i=1}^N (\zeta_i + \zeta_i^*)$$

Subject to :

$$y_i - w^T \phi(x_i) - b \leq \epsilon + \zeta_i$$

$$w^T \phi(x_i) + b - y_i \leq \epsilon + \zeta_i^*$$

$$\zeta_i, \zeta_i^* \geq 0$$

Where

$w$  is the weight vector

$b$  is the bias term

$\phi(x_i)$  is the feature map (often nonlinear transformations of  $x_i$ )

$\zeta_i$  and  $\zeta_i^*$  are slack variables that allow for some degree of error

Dual Formulation:

SVR can also be formulated in its dual form for computational efficiency:

$$\max_{\alpha, \alpha^*} \sum_{i=1}^N (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j=1}^N (\alpha_i^* - \alpha_i)(x_i \cdot x_j)$$

Subject to :

$$0 \leq \alpha_i^*, \alpha_i \leq C$$

$$\sum_{i=1}^N (\alpha_i^* - \alpha_i) y_i = 0$$

Where  $\alpha$  and  $\alpha^*$  are Lagrange multipliers associated with the constraints

#### 4. RESULTS&DISCUSSION:

##### 4.1 Raman spectra:

Figure 2 presents the Raman spectra of the glasses under investigation, showing structural changes when Bi<sub>2</sub>O<sub>3</sub> is replaced with La<sub>2</sub>O<sub>3</sub>. Small peaks around 395-401 cm<sup>-1</sup> relate to Zn-O bond stretching in ZnO [9, 12], and peaks around 924 cm<sup>-1</sup> correspond to other ZnO vibrational modes. Peaks around 501-508 cm<sup>-1</sup> are associated with Zr-O bond stretching in ZrO<sub>2</sub> [13], while peaks around 1033 cm<sup>-1</sup> may correspond to other ZrO<sub>2</sub> vibrational modes [11]. A strong peak around 600 cm<sup>-1</sup> is linked to the conversion of BO<sub>3</sub> to BO<sub>4</sub> units in B<sub>2</sub>O<sub>3</sub> [9, 14], with intensity increasing and shifting to higher wavenumbers when La<sub>2</sub>O<sub>3</sub> is added, confirming structural changes. Peaks around 1164-1210 cm<sup>-1</sup> correspond to other borate group vibrations [9, 15], decreasing in intensity and shifting to lower wavenumbers with increasing La<sub>2</sub>O<sub>3</sub> content. Peaks around 817-824 cm<sup>-1</sup> are assigned to Si-O bond stretching in SiO<sub>2</sub> [16], characteristic of silica-based glasses. No significant peaks for Na<sub>2</sub>O were observed, as it acts as a network modifier rather than forming strong bonds that contribute to Raman peaks. Raman spectra prediction using Regression learning techniques

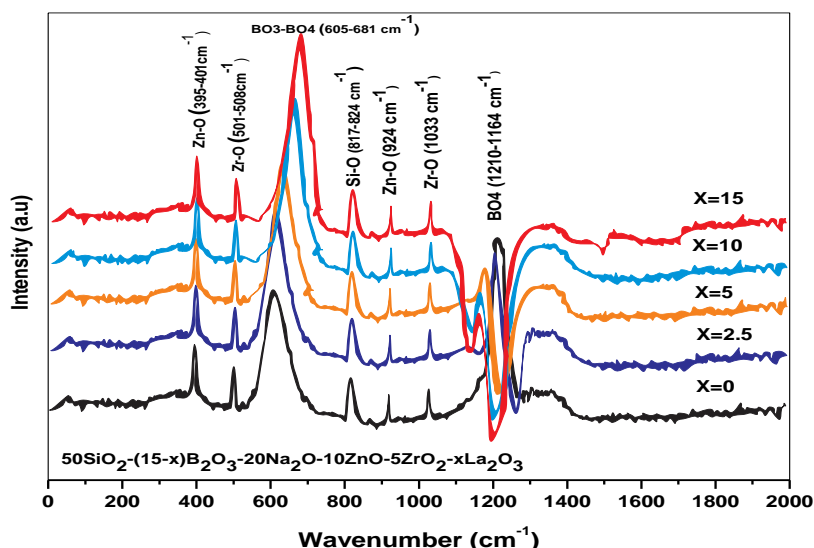


Figure 2 Raman spectra of Present glass samples

##### 4.2 Support Vector Machine (SVM):

SVM is a supervised regression tool that can be used for both classification and regression problems. SVM for classification tasks using scikit-learn library in Python. The prediction of Raman spectra using SVM as follows. The experimental Raman spectra data stored of various borosilicate glasses in variables x (wavenumber) and y (intensity), the code proceeds to split this data into training and testing sets using the train\_test\_split function. The test size considers as 20% i.e 0.2 of total data which is used for testing of the algorithm, other remaining 80% will be used for training the algorithm. A random\_stae is set to 42 for best prediction of the current dataset. The features (X\_train and X\_test) are standardized to have a mean of 0 and a standard deviation of 1 using the StandardScaler class. An SVM classifier (svm\_classifier) is initialized with specified parameters such as the kernel type (kernel='rbf'), regularization parameter (C=1.0), and gamma value (gamma='scale'). The Radial Basis Function (RBF) kernel is chosen as it is suitable for handling non-linear decision boundaries. The SVM classifier trained on the standardized training data (x\_train\_scaled and y\_train) using the fit method. The trained SVM classifier is used to predict labels for the test set (x\_test\_scaled) using the predict method, resulting in predicted labels (y\_pred). The classification report function provides various parameters such as recall, F1score, precision and support for each test class.

Precision is the ratio of correctly predicted positive observations to the total predicted positives. In the output, the present code got 0.969 for all classes. This means that for each class, the model successfully make correct positive

predictions out of the total predicted positives. Recall, also known as sensitivity, is the ratio of correctly predicted positive observations to the all observations in actual class. Similar to precision, it's displayed as 0.975 for all classes, indicating that the model correctly predict positives out of the actual positives for each class. The F1-score, representing the harmonic mean of precision and recall, achieves a balance between the two metrics. In the current algorithm, the F1-score is 0.997, indicating that both precision and recall are nearly perfect for each class. Support refers to the number of actual instances of each class within the specified dataset. In the output, it's a constant value of 0.99 for each class, indicating that each class appears only once in the dataset. Accuracy is the proportion of correctly classified instances among the total instances.

The correlation heatmap in the provided code visualizes the pairwise correlation coefficients between the spectral features in the Raman spectra dataset, highlighting the relationships among them. Using `sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f")`, the heatmap shows how each pair of features correlates, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Figure 3 present the correlation heatmap which shows the precision, recall F1-score, and support. Figure 4 presents the experimental verses predicted Raman intensity of the test data. This plot show the data is consistent with experimental values. Figure 5 present the Raman spectra generated from the SVM learning algorithm. The generated spectra similar to the experimental Raman spectra. Table 1 present the Raman band assignments from both experimental and ML generated. It was also observed from the spectrum that the band assignments similar to the experimental assignments and the band related to the BO3 and BO4 units shifting lower and higher wavelengths as present in experimental spectrum. Therefore the SVM algorithm successfully predicting the Raman spectra of present glasses.

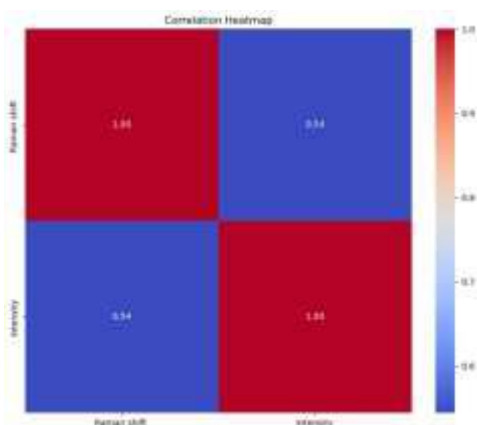


Figure 3 correlation heatmap of SVM

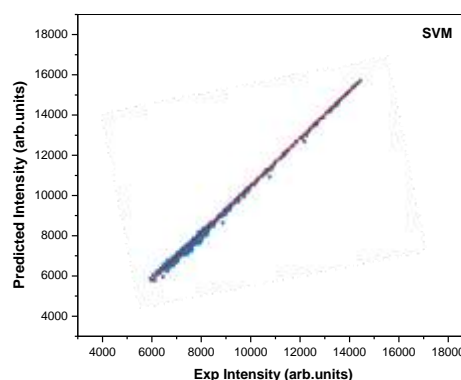


Figure 4 experimental verses predicted Raman intensity of the test data

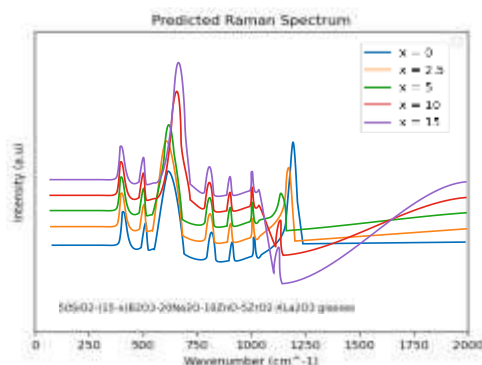


Figure 5 Raman spectra generated from the SVM learning algorithm

### 4.3 Random Forest Regression

Random Forest Classifiers are ensemble algorithms constructing multiple decision trees, outputting mode of classes for predictions [8,9]. Key steps involve importing libraries for visualization, ML, pandas for data processing, scikit-learn for ML tasks like data splitting, feature scaling, training classifiers, and performance evaluation. Raman spectral data imports from Excel into a pandas DataFrame, undergoing preprocessing like splitting into wavenumber (X) and intensity (y), and visualizing feature correlations with a heatmap. Data splits into training/testing sets with a

specified test size (e.g., 20%) and standardized features using StandardScaler. Evaluation includes precision, recall, F1-score metrics, possibly saving the model. Precision is 0.958, recall is 0.967, and the F1-score is 0.988 for all classes, indicating high accuracy. Each class appears once, with support at 0.97. Accuracy represents the proportion of correctly classified instances. Figure 6 presents the fitting random regression with various random states. Figure 7 presents the experimental verses predicted Raman intensity of the test data. This plot show the data is consistent with experimental values and random forest classifier is successfully predicting the experimental data. Table 1 provides the Raman intensity predicted values obtained from the Random forest. Table 1 clears the experimental data consist with predicted data.

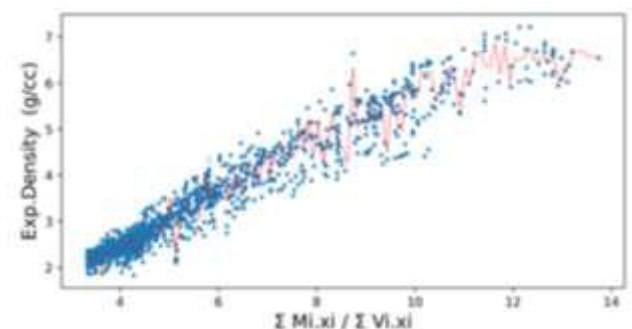


Figure 6 Random forest regression fitting

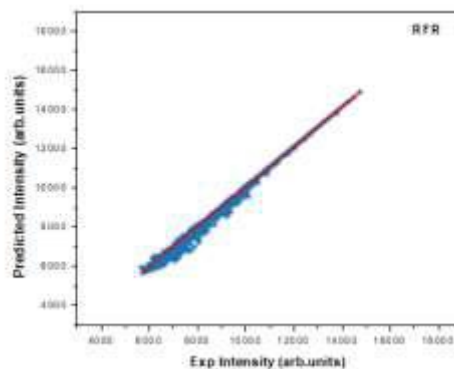


Figure 7 Predicted band assignments plot

Experimental Raman Assignments		Predicted Raman Assignments	
Wavenumber (cm <sup>-1</sup> )	Band Assignments	SVM	Random forest
		Wavenumber (cm <sup>-1</sup> )	
395-401	Stretching vibrations of Zn-O bonds	393-404	392-400
501-508	stretching vibrations of Zr-O bonds	499-512	497-504
605-681	Conversion of BO <sub>3</sub> to BO <sub>4</sub> units of BO <sub>3</sub>	612-661	614-660
817-824	symmetric stretching vibrations of Si-O bonds	810-820	811-818
924	vibrational modes of ZnO	970	965
1033	vibrational modes of zirconia	1013	1014
1164-1210	stretching vibrations BO <sub>3</sub> units	1131-1200	1130-1205

Table 1 Raman band assignments predicted parameters

## 5. CONCLUSION:

The present study has demonstrated the successful synthesis and characterization of glasses with the composition 50SiO<sub>2</sub>-(15-x)B<sub>2</sub>O<sub>3</sub>-20Na<sub>2</sub>O-10ZnO-5ZrO<sub>2</sub>.xLa<sub>2</sub>O<sub>3</sub> (where x ranges from 0 to 15) using the melt-quenching method. The following conclusions can be drawn from the study. Structural changes in the Raman spectra were observed, particularly the conversion of BO<sub>3</sub> to BO<sub>4</sub> units and shifts in peak intensities and positions, indicating modifications in the glass network due to varying La<sub>2</sub>O<sub>3</sub> content. SVM successfully predicted Raman spectra and achieved high precision (0.969), recall (0.975), and F1-score (0.997) and confirming the effectiveness of SVM in capturing spectral features and trends. Random forest classifiers achieved high precision (0.969), recall (0.975), and F1-score (0.997) in predicting Raman spectra and produced consistent predictions aligning closely with experimental Raman spectra

**REFERENCES :**

1. M.Y. Hassaan, S.A. El-Badry, M. Tokunaga, T. Nishida, , Materials Letters, 59(2005), 3788-3790.
2. Gurbinder Kaur, O.P. Pandey, K. Singh, Journal of Non-Crystalline Solids, 358 (2012) 2589-2596.
3. O.I. Sallam, A.M. Madbouly, F.M. Ezz-Eldin, Journal of Non-Crystalline Solids, 590 (2022), 121691.
4. N. Hadj Youssef, M.S. Belkhiria, J.J. Videau, M. Ben Amara, Materials Letters, 44(2000), 269-274.
5. Ufuoma Joseph Udi , Mustafasanie M. Yussof, Kabiru Musa Ayagi , Chiara Bedon, Mohd Khairul Kamarudin, Ain Shams Engineering Journal, 14 (2023) 101970
6. Ravindranadh Bobbili, Materials Letters, 349(2023), 134774.
7. Ravindranadh Bobbili, B. Ramakrishna, Materials Today Communication, 36(2023)106674.
8. Binghui Deng, Journal of Non-Crystalline Solids, 529 (2020) 119768.
9. Shaik Amer Ahmed, Shaik Rajiya, M. A. Samee, Shaik Kareem Ahmmad, Kaleem Ahmed Jaleeli, Journal of Inorganic and Organometallic Polymers and Materials, 32(2022) 941-953.
10. Shaik Kareem Ahmmad, Nameera Jabeen, Syed Taqi Uddin Ahmed , Shaik
11. Amer Ahmed, Syed Rahman, Ceramics International, 47(2021)7946-7956
12. Shaik Kareem Ahmmad, Norah A.M. Alsaif , M.S. Shams, Adel M. El Refaey, R.A. Elsad, Y.S. Rammah, M.S. Sadeq, Optical Materials 134( 2022) 113145
13. S G MOTKE , S P YAWALE and S S YAWALE,Bull. Mater. Sci., 25 ( 2002)75–78
14. S. W. Lee, R. A. Condrate Sr , *J Mater Sci* 23 (1988) 2951–2959.
15. V Sudarsan, V K Shrikhande, G P Kothiyal and S K Kulshreshtha, *J. Phys.: Condens. Matter* 14
16. Chandkiram Gautam,Avadhesh Kumar Yadav,and Arbind Kumar Singh Volume 2012, Article ID 428497, 17
17. Nur Ramadhan Mohamad Azaludin , Nurul Syahidah Sabri, gading Journal of Science and Technology, Vol 4 No (1) (2021)