CHAPTER FOUR

RESULTS AND DISCUSSION

In this chapter, the performance of the new models are compared using different metrics of accuracy. The accuracy of each model is first tested, followed by a comparison between the models and the base case (Standing and Katz chart) and finally comparison of the best model with other correlations considered in this chapter. The results obtained from the first two sections help us to understand the predictive power of machine learning models and select the best model which was compared with other relevant correlations.

4.1 MODEL ACCURACY

4.1.1 SVR ACCURACY

As explained in the previous chapter, in building this support vector regression model, a grid search was used which allows for the combination of different hyper-parameters so as to select the best combination in terms of some metric(s). This is what is commonly referred to as hyper-parameter optimization. The parameters of SVR considered were 'C', and 'gamma'. These are the two most important parameters to consider when building an SVR model (Vanderplas, 2017). The resulting model at the end of the training phase had a C and gamma value of 20 and 1 respectively. SVR(C=20, gamma=1). Although the SVR can be a very good estimator with the default hyper-parameters, tuning becomes important to achieve better results thereby bringing the best out of the model. Table 1 shows the difference before and after optimization and also validates the need for this process.

Table 4. 1: Table of accuracy for the SVR model

	SVR before	SVR after	Difference
Train R^2 score	0.94895	0.98222	0.03352
Test R^2 score	0.95836	0.98272	0.02406
MAE	0.07359	0.06019	-0.013
MSE	0.01127	0.00476	-0.00651

RMSE	0.10616	0.06903	-0.03718

4.1.2 GBDT ACCURACY

This model with default parameters performed better than the SVR model based on all the metrics used for evaluation. This shows how powerful this ensemble model can be. A grid search was also done in building this model to get the best result. Optimization becomes important here to improve accuracy and avoid overfitting. Table 2 shows the effect of optimizing this model because the model tends to overfit easily and needs to be flexible because of the nature of the isotherms in the SK chart. It shows that the GBRT gives the best correlation coefficient accuracy and minimum error.

Table 4. 2: Table of accuracy for the GBDT model

	GBDT before	GBDT after	Difference
Train R^2 score	0.99787	0.99996	0.00212
Test R^2 score	0.99674	0.99962	0.00287
MAE	0.02161	0.00561	-0.01615
MSE	0.00088	0.00010	-0.00077
RMSE	0.02966	0.01033	-0.01933

4.1.3 RBF-NN ACCURACY

The RBF-NN model was built by manually adjusting some of the most influential parameters which are the optimizer and learning rate (Goodfellow, Bengio, & Courville, 2016). The optimizer for a neural network is responsible for reducing the loss function thereby increasing performance. The learning rate for an optimizer controls how fast the error loss is updated. The optimizer used was RMSprop and a learning rate of 0.01 was considered the best.

The built RBF-NN model gave a training R^2 score, test R^2 score, MAE, MSE and RMSE of 0.97723, 0.97467, 0.05473, 0.00663 and 0.08144 respectively.

4.2 COMPARISON BETWEEN THE ACTUAL AND PREDICTED VALUES

The word 'actual' will be used from time to time to refer to the basis of comparing the models. Figure 4.1 shows the frequencies of the actual compressibility values (histogram). From this distribution it can be seen that most frequent values of z falls within 1.5 to 1.75. The next most frequent values fall within 0.75 to 1.0. This figure serves as the basis for comparison with the three built models as shown in figure 4.2 – 4.4. The distribution of z-factor predicted by the gradient boosted regression model in figure 4.2 shows a close similarity with that of figure 4.1. The two most frequent range of z-factor for the GBDT model is the same for the actual and also a general look at the two reveals the closeness of the two distributions. Figure 4.3 shows the distribution obtained with the RBF-NN. Some bars in Figure 4.3 shows an over-estimation when compared to the actual. The inaccuracy in properly estimating the z-factor can be attributed to the fact that neural networks require much more data than the traditional machine learning algorithm (Geron, 2019).

For the last distribution, Figure 4.4 shows the distribution of predicted z-factor obtained from SVR model. This distribution shows a closer relation to the actual than the RBF-NN model. This shows that although the model is simple, it can make optimal predictions with the right set of hyper-parameters. The difference between the two distributions might not be obvious at first glance. Taking a general look at the three distributions gotten from the models, the models can be ranked by order of accuracy (closeness to the actual z-factor data distribution).

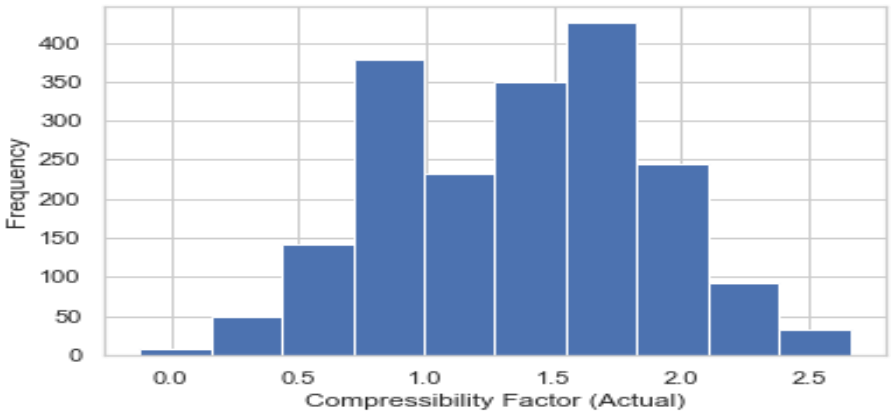

Figure 4.1: Histogram of actual z-factor

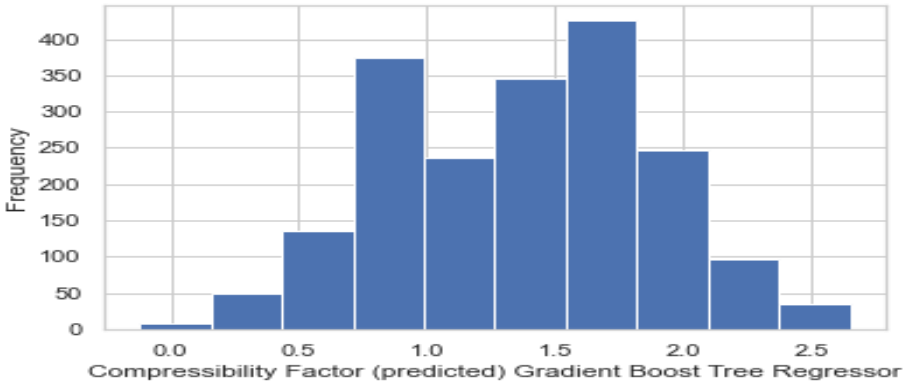

Figure 4.2: Histogram of predicted z-factor for GBDT

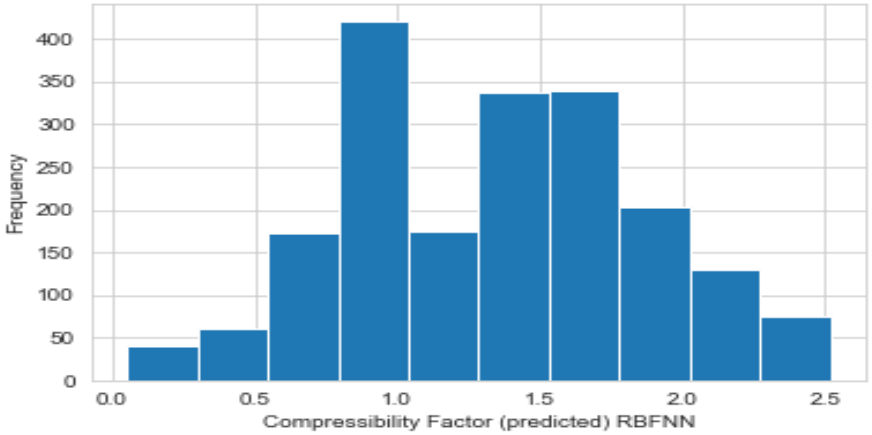

Figure 4.3: Histogram of predicted z-factor for RBFNN

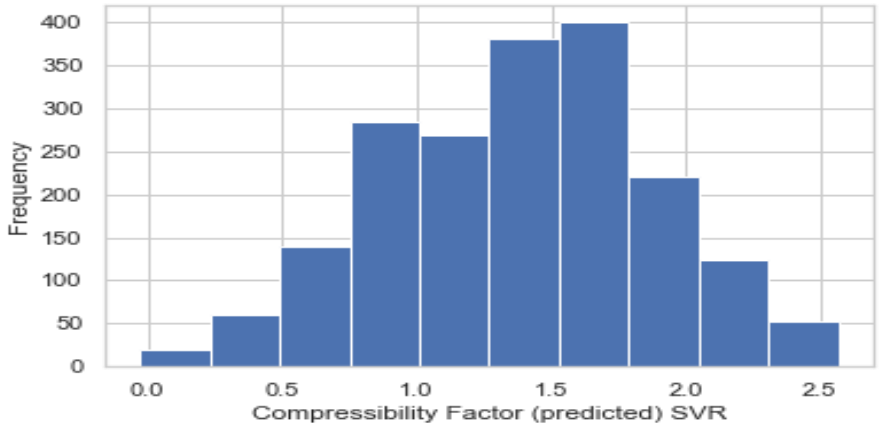

Figure 4.4: Histogram of predicted z-factor for SVR

The R^2 score or the coefficient of determination for all three built models showed very good results. A further look at regression plot verifies the accuracy of the gradient boosted regression model over the other models. In Figure 4.5, it can be observed that the RBFNN model had predicted values that deviated far from the actual at lower and the upper portion along the line. The SVR performs better than the RBFNN model in figure 4.6, given that the points are relatively closer to the black line. The GBDT model performed the best given that almost all data points fall under the black line as seen in figure 4.7.

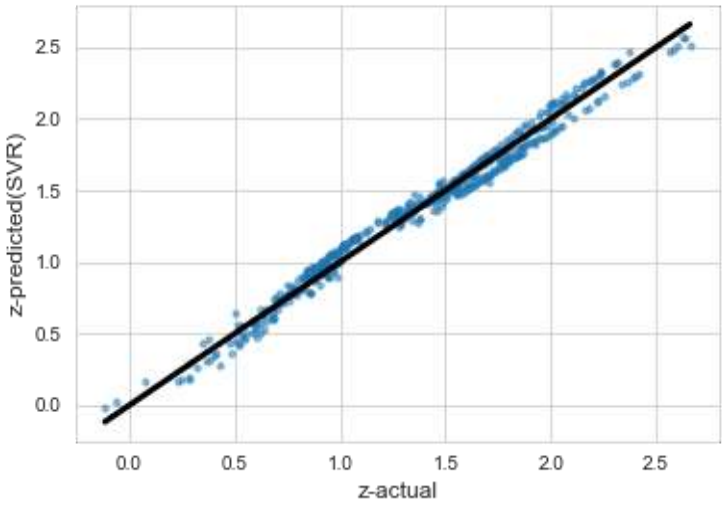

Figure 4.5: Regression plot for SVR

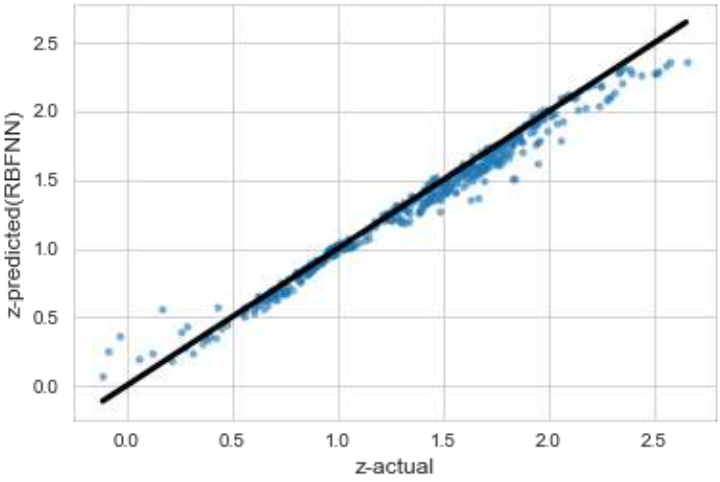

Figure 4.6: Regression plot for RBFNN

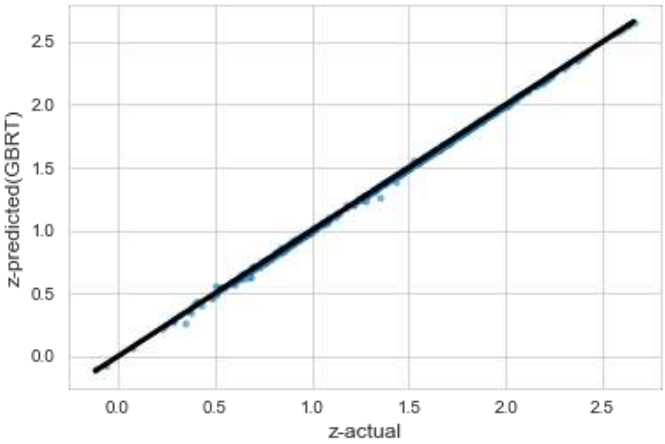

Figure 4. 7: Regression plot for GBDT

Figure 4.8-4.12 shows a plot of the compressibility factor against pseudo-reduced pressure at certain pseudo-reduced temperature of 0.92, 0.95, 1, 1.5, 2, 2.4 and 3.0.

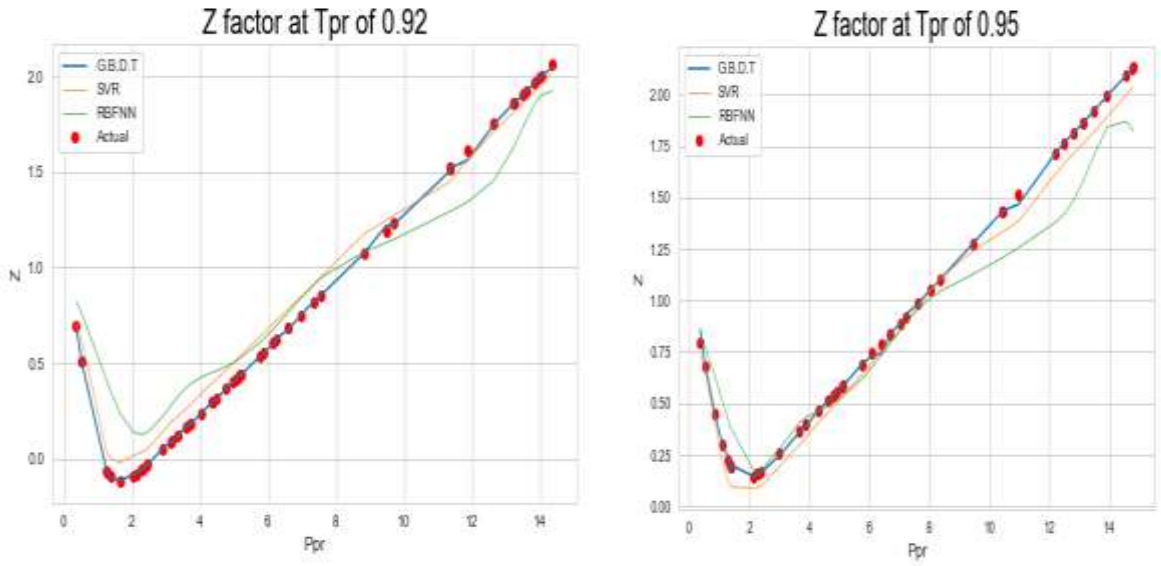

Figure 4. 8: isotherm of 0.92 and 0.95

From figure 4.8, the GBRT almost fits the actual data points perfectly, so therefore for this model it is expected that the predictions at the isotherm 0.92 is somewhat accurate. The 'actual' refers to the data from beggs and brill which covers low temperature and standing and Katz for moderate to higher temperatures. The SVR follows the trend but tends to over-estimate for almost all P_{pr} points at Tpr of 0.92. The RBFNN has the largest deviation from the actual. The next plot in figure 4.9, shows that the models become a little bit less accurate than for the first isotherm considered. This can be observed from the increased deviation towards the upper end or high pseudo reduced pressure.

For the Tpr at 1.0 and 1.5, the GBRT model seems to still be matching the actual curve with little errors.

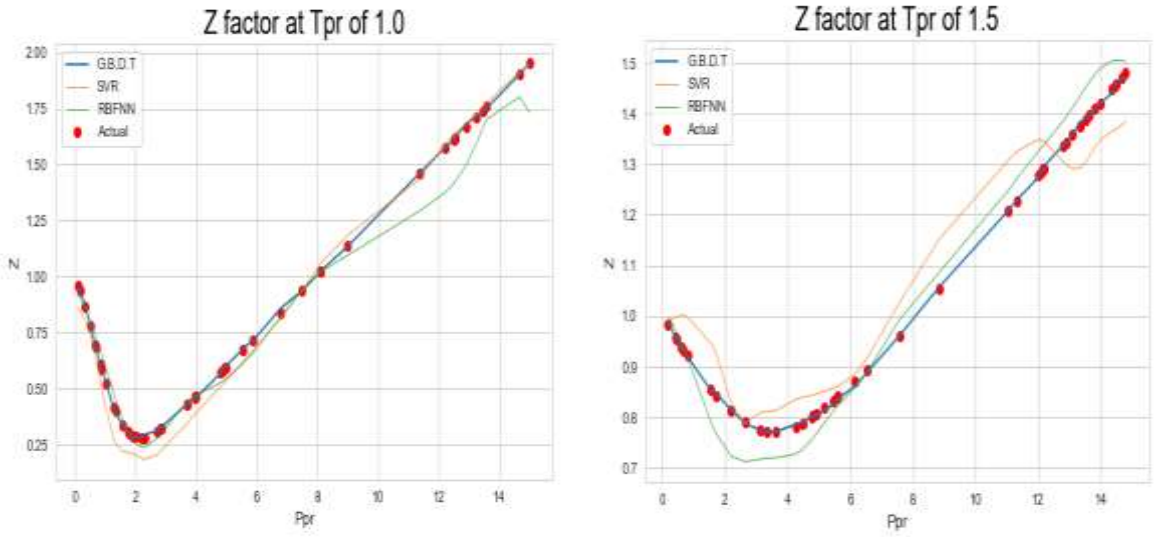

Figure 4. 9: Isotherm of 1.0 and 1.5 predicted

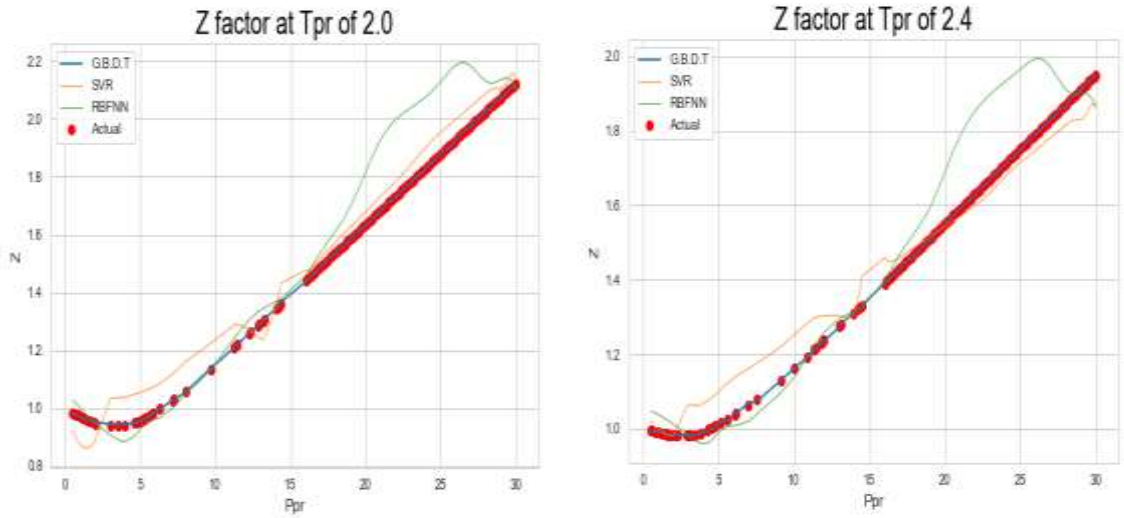

Figure 4.10: Isotherm of 2.0 and 2.4 predicted

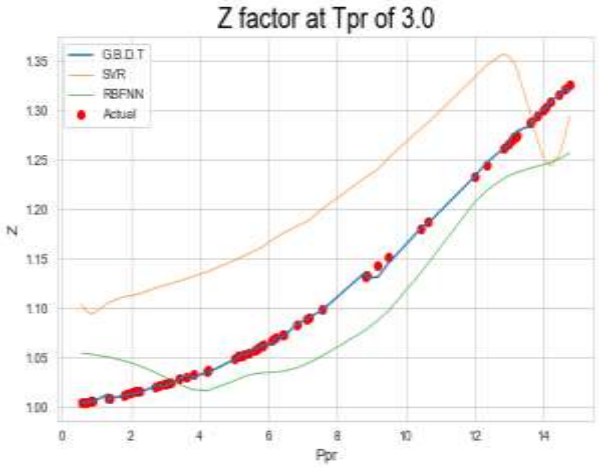

Figure 4. 11: Isotherm of 3.0 predicted

At higher Tpr values in figure 4.6 to 4.8, the SVR model and RBFNN model seems to be unable to make accurate predictions. The GBDT model continues to prove its superiority in making prediction by its smooth match to the actual curve from figures 4.6 to 4.8.

For all the isotherms considered, the GBDT model had a great fit. The SVR model was better at lower isotherms than higher isotherms. And finally the RBFNN had the worst fit progressing from the lower isotherms up to the highest isotherm considered.

4.3 COMPARISON WITH OTHER CORRELATIONS

Table 4.3 shows the results of the prediction of z-factor at different T_{pr} and P_{pr} values by correlations developed in previous studies and in this study. Beggs and Brill, Orodu et al., sanjari and Lay, Ekechukwu and Orodu studies were compared to the proposed model. The benchmark for this comparison was the standing and Katz chart in the fourth column of Table 4.3. The Orodu et al. correlation was selected because it was developed for low T_{pr} and as well as the Beggs and Brill correlation. At most of the selected pair of T_{pr} and P_{pr}, the GBDT model in general was seen to have predicted better than the other correlations. Based on the points from the table, the mean absolute percentage error was calculated and plotted as seen in *Figure 4.12.* From the plot, it is clear that the proposed model outperformed the other correlations and therefore verifies the authenticity of this model.

Table 4.3: Z-factor correlation values at various P_{pr} and T_{pr}

T_{pr}	P_{pr}	Beggs and Brill	Standing and Katz	Orodu et al.(2019) 1	Orodu et al.(2019) 2	Sanjari & Lay, (2012b)	Ekechukwu & Orodu, (2019)	Model1(GBDT)
1.35	0.2	0.976	0.97	1.217	0.735	1.002	0.976	0.974
1	1	0.862	--	0.967	0.757	1.011	0.429	0.527
1.15	2	0.739	0.465	0.72	0.757	1.124	0.473	0.459
1.2	3	0.671	0.535	0.598	0.745	0.442	0.544	0.542
1.25	4	0.674	0.63	0.597	0.751	0.559	0.643	0.632
1.3	5	0.738	0.718	0.666	0.809	0.671	0.737	0.733
1.35	6	0.816	0.815	0.757	0.875	0.775	0.825	0.819
1.4	7	0.894	0.9	0.852	0.937	0.869	0.907	0.894
1.45	8	0.972	1	0.95	0.995	0.954	0.984	0.99
1.5	9	1.05	1.08	1.05	1.0496	1.031	1.057	1.065
1.6	10	1.127	1.135	1.151	1.102	1.103	1.127	1.137
1.7	11	1.203	1.2	1.254	1.153	1.164	1.191	1.203
1.8	12	1.28	1.25	1.358	1.201	1.218	1.251	1.244
1.9	13	1.356	1.3	1.464	1.248	1.268	1.307	1.299
2	14	1.432	1.34	1.571	1.293	1.315	1.358	1.344
2.2	15	1.508	1.36	1.679	1.336	1.358	1.402	1.356

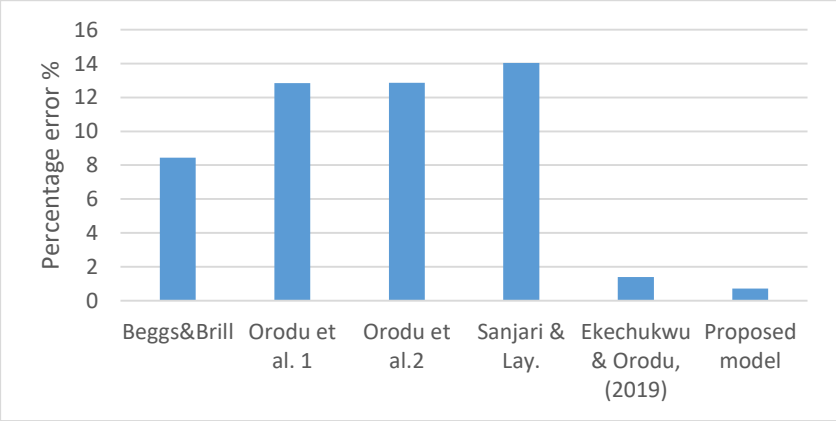

Figure 4. 12: Mean Absolute Percentage error for correlation

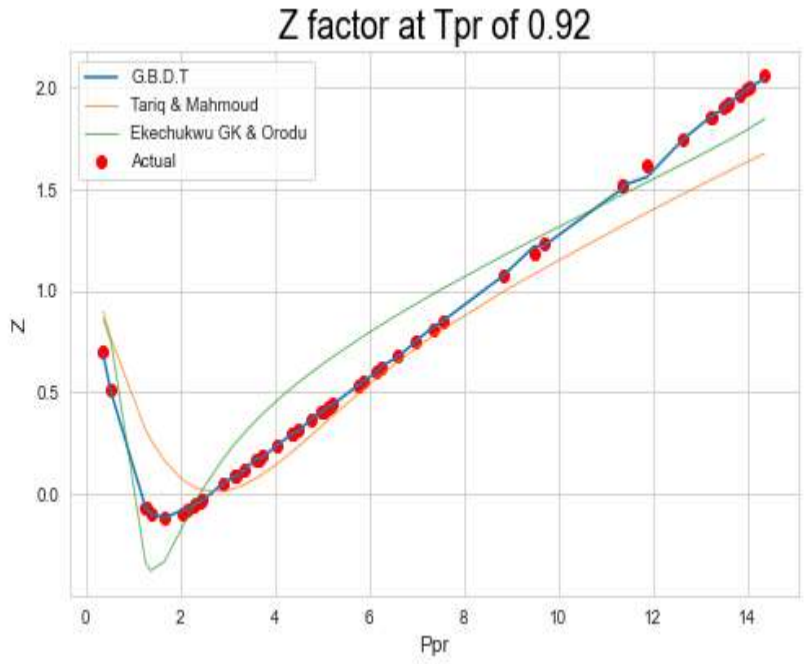

Figure 4.13: Z-factor prediction for notable correlation at Tpr 0.92

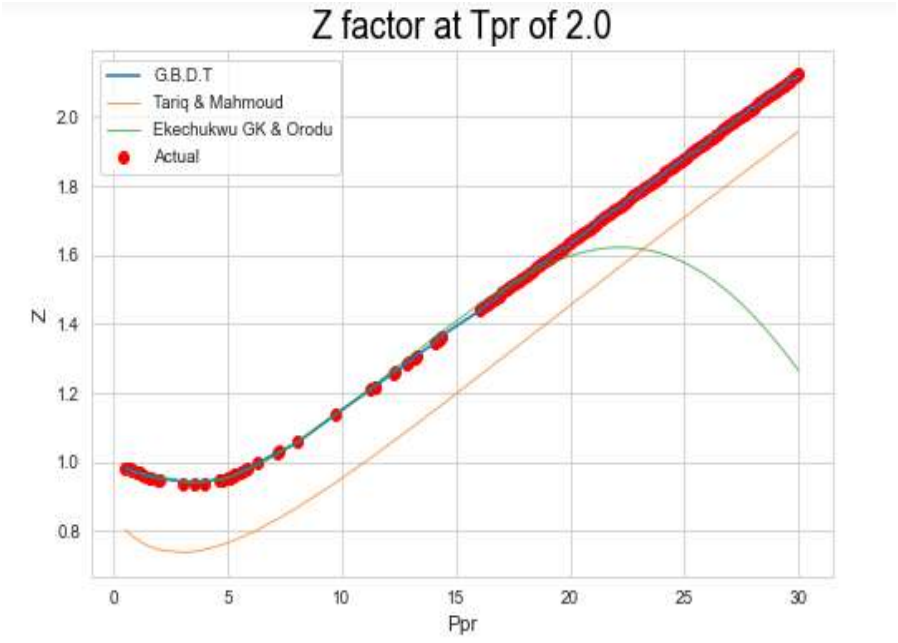

Figure 4.14: Z-factor prediction for notable correlation at Tpr 2.0

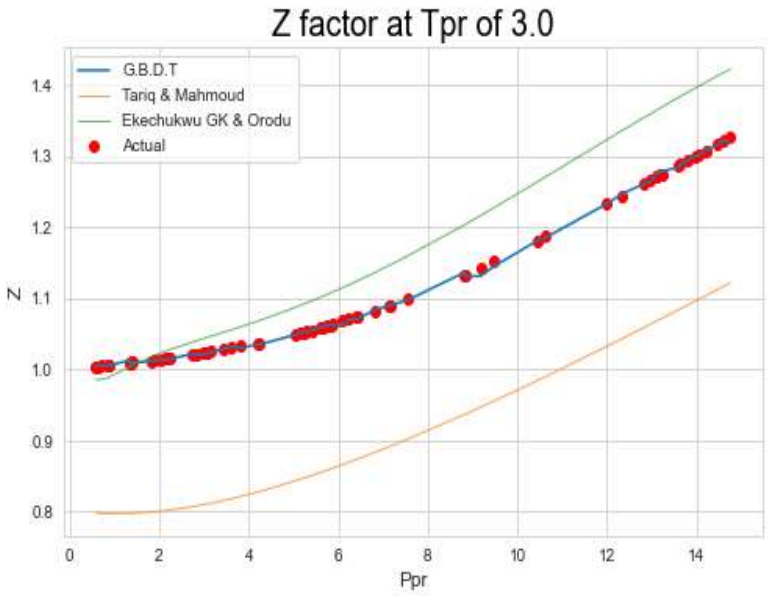

Figure 4.15: Z-factor prediction for notable correlation at Tpr 3.0

The selected isotherms for the plot from figure 4.13 to 4.15 were chosen to show the important areas of the model. Tariq & Mahmoud (2019) and Ekechukwu & Orodu(2019) models was selected because they have wide range of applicability so therefore will render the proposed model somewhat valid if it is observed to perform better. In figure 4.13, it shows that at low T_{pr} the GBDT model is able to match the data points of the actual more than the rest. In figure 4.14, the behavior of the models at T_{pr} of 2.0 and wide range of P_{pr} up to 30 is shown. Ekechukwu & Orodu(2019) model performed well at lower P_{pr} values but did not do so well with P_{pr} values beyond 18. Tariq & Mahmoud model matches the pattern or curve but under-predicts the values. For figure 4.15, it shows that even at high T_{pr} the proposed model continues to match the data points accurately. Ekechukwu & Orodu(2019) tends to over predict the z values while the Tariq & Mahmoud (2019) model still under estimates the values of z-factor.