Prediction of Pressures as a Result of Air & Liquid Flow

By Group 2 (Lisa Reisenauer, Triton Wolfe, Joshua Randrup, Zhenzi Yu, Yuhan Yang)

PROJECT BACKGROUND

Anticipating phenomena is essential to the smooth and accurate functioning of a production environment. As such, a significant amount of resource is invested into the modeling and prediction of various phenomena that can affect process reliability and variability. Some examples include fouling factors in heat exchangers, inefficiencies in pumps due to varying Reynolds number, or whirls and eddies that can occur in flow reactors. This project looks to model the behavior of two-phase flow such as that which is present in the production of Methacrylic Acid as patented by Rohm and Haas. This process injects air into a liquid stream as a reactant at several stages. The resulting two-phase flow has a behavior which is extremely difficult to predict. Theoretical mathematical models of two-phase flow employ a categorical separation of types of two-phase flow into one of the following: Transient flow, Separated flow, or Dispersed flow. Each of the categories then has different models to predict their behavior. Unfortunately, that math can be extremely variable based on gas behavior assumptions. As such, they cannot be assumed to be good models for real world use. This project aims to take behavior from a real-world two-phase flow and create a model to reliably and accurately predict various pieces of the system. This procedure could then be retrained using production environment data. The retrained model would then produce reliable and accurate predictions of the process' event's and behavior.

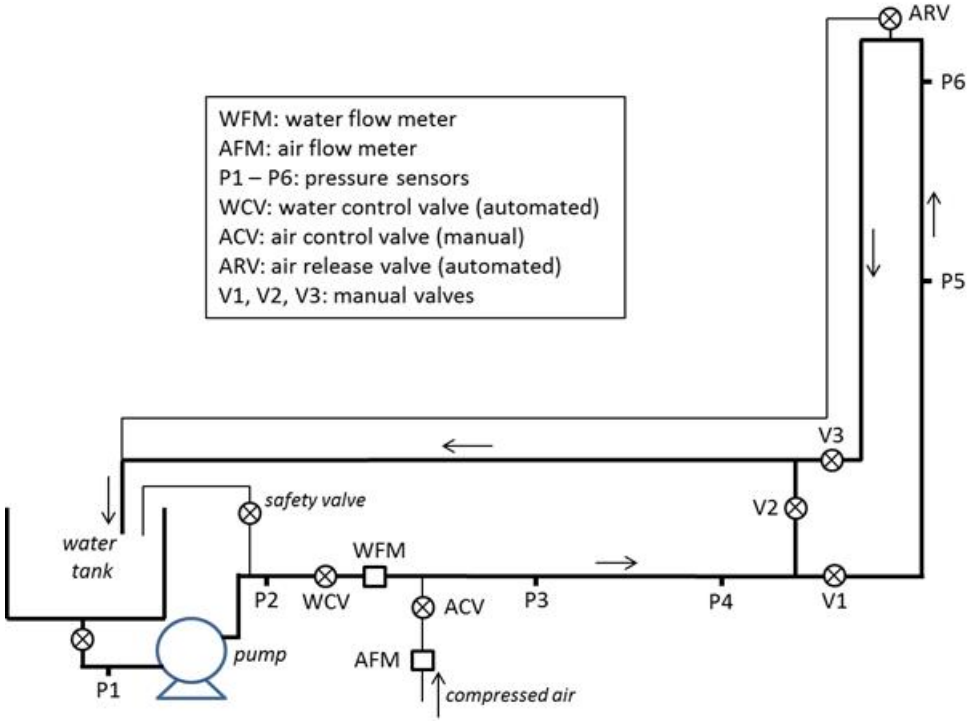

Figure 1. Schematic diagram of two-phase flow

PROJECT OBJECTIVE & GENERAL STRATEGY

A unit operations set-up on Georgia Tech's campus injects air into a liquid water stream and measures downstream pressures. The schematic diagram is shown as Fig. 1. This project aims to use data from this set-up to create a regression model that can correlate the downstream pressures to the pressures in upstream and the air/water flow rates. Namely, the single target variable is P6 (Fig. 1), features are P1-P5, air/water flow rates.

After data exploratory analysis and cleaning, algorithms that have low complexity will be considered first to build a baseline model. The baseline model can later be optimized by utilizing time-series analysis and modeling techniques. Principle component analysis and feature engineering will also be implemented if necessary, to further improve the model.

WORKFLOW

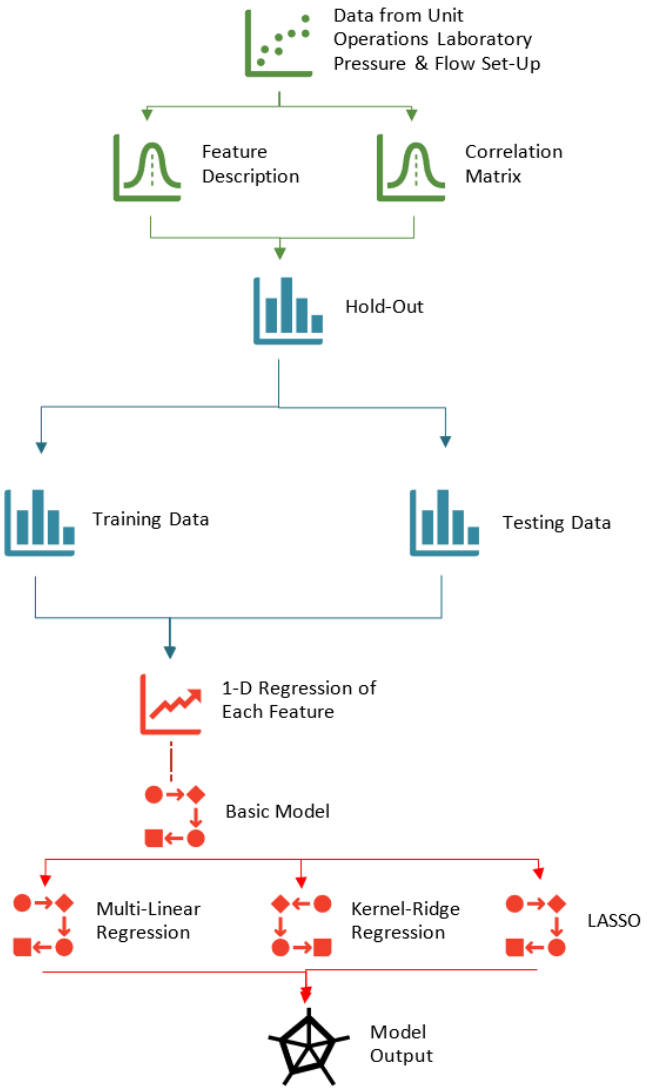

Figure 5. Workflow of the project

DATA EXPLORATORY ANALYSIS

Feature Distribution

The dataset, provided by the Georgia Tech ChBE Unit Control Lab, was obtained during the two-phase flow correlation experiment. It consists of 8 parameters: air flow rate, water flow rate and 6 pressures along the pipeline. We will use the pressure-6 as our target variable, and the remaining 7 features as inputs. We have 8057 data points available, which is a good sample for this system, considering the number of features.

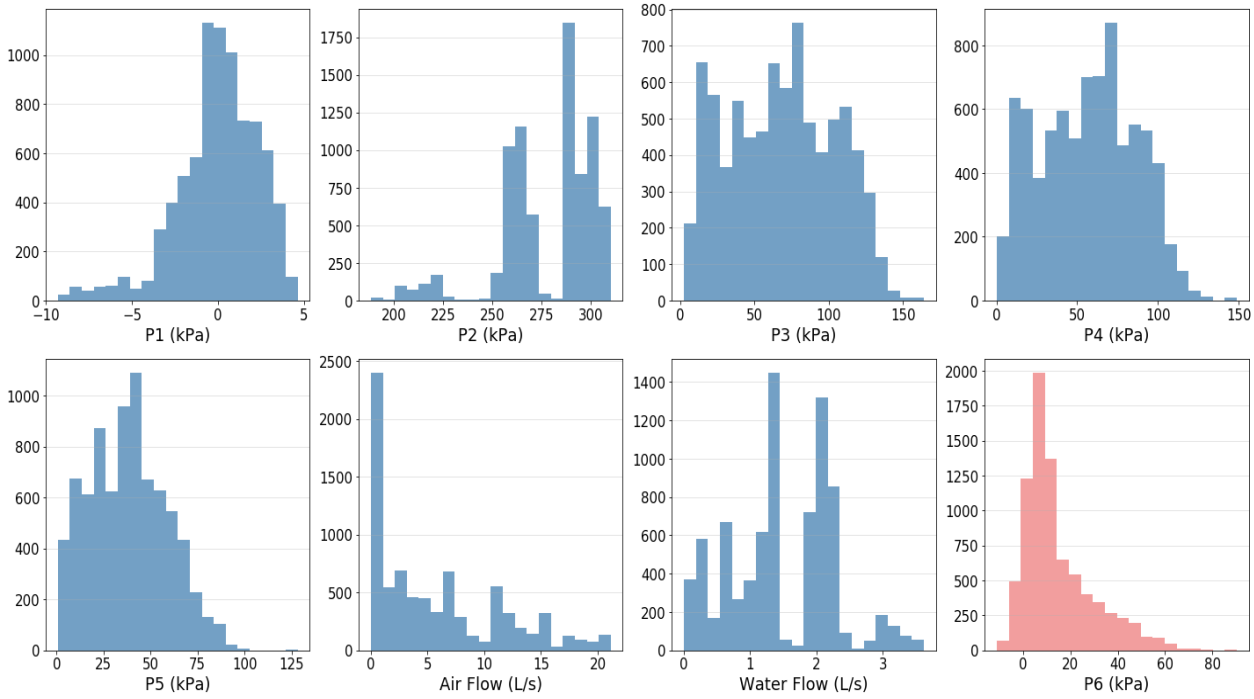

Figure 2. Distribution of the variables. Features are colored bule and the target variable is colored coral.

As shown in Fig. 2, after data visualization, there are mainly two types of distribution: near normal and random. Moreover, the orders of magnitude of these features vary considerably. Thus, we will rescale the features before further processing.

Correlation Matrix

To further study the relationships between the features, we calculated the correlation matrix of this system. As shown in Fig. 3, P3, P4 and P5 are highly correlated (>0.9) while P2 and water flow rate are anti-correlated (<-0.9). These observations are consistent with the physics of the physical system. We could further reduce the feature space's dimensions based on the correlation matrix, however the total number of the features is not large and computational workload is not heavy in this system, we choose to keep all these features to maximize the prediction accuracy.

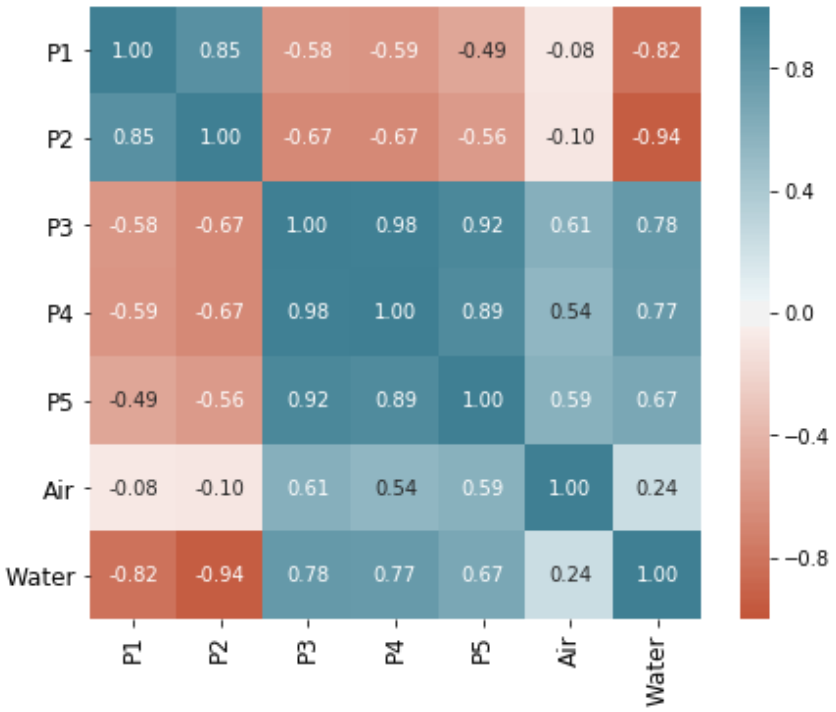

Figure 3. Heatmap of correlation between scaled feature values, color-coded from-1.0 (coral) to 1.0 (blue)

Single component regression

We also did a 1-D regression to study the correlations between the target variable (P6) and each single feature using a piecewise function and a rbf kernel function (Tab.1). The piecewise linear model gave a good R-score on the training set, but not on the testing set. These results make sense since the non-parametric models are not good at extrapolation.

Table 1. R-scores of the single component regression (RBF Kernel values are those of the highest training R-scores found during parameter tuning.)

	Piecewise		RBF kernel	
	training	**test**	**training**	**test**
P1	0.936	-2.017E3	0.926	-9.570
P2	0.991	0.582	0.990	-1.325E14
P3	0.995	0.490	0.995	-1.992
P4	0.998	0.451	0.998	-8.700
P5	0.996	0.490	0.996	-6.138
Air	0.909	0.496	0.909	-2.561E13
Water	0.813	0.545	0.801	-1.846E15

The RBF kernel was also found to be an inadequate model for predicting the downstream pressure based on any single feature. When tuned to an optimal gamma (Fig.4), the R-scores on the training set were nearly 1. However, the testing scores were consistently negative, indicating that the error from the model was greater than the error would have been if the mean of the features' data had been used as the model itself.

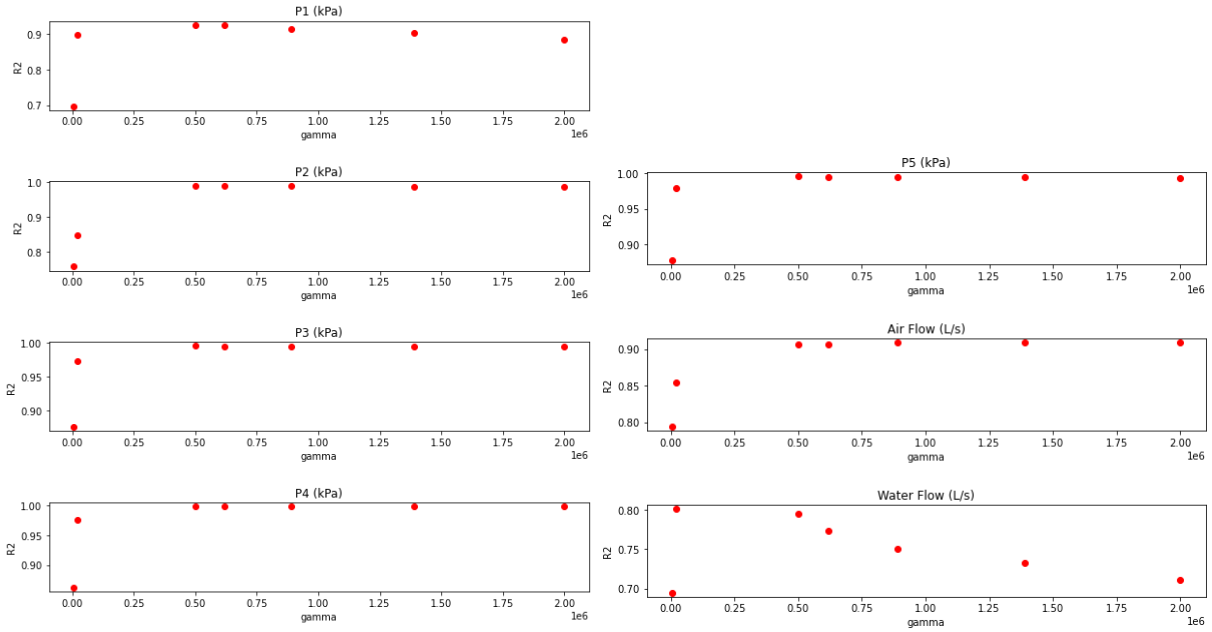

Figure 4. R score of each single component regression as a function of gamma

BASELINE MODEL

The baseline model pipeline is shown as Fig.5. As a starting point for analyzing our data and determination of likely favorable models, a correlation matrix was created for the various features of our data. Variables P3, P4 and P5 show a relatively linear relationship; the 1-D regression analysis indicates a single feature is not enough to accurately predict our target variable (P6).

Linear, non-linear, non-parametric models and time-series models were implemented at this exploratory stage. Specifically, the models implemented were multi-linear regression, Kernel Ridge Regression, Lasso, and time-series. As shown in Fig.6, the multi-linear model got comparable r2 score for training set and test set. But the model tends to underestimate the P6 in the high-pressure region. It is not surprising that the multi-linear model did not work well, since it is the most basic form. The low R score shows the model underfitted data, indicating the number of features is relatively small or a linear model is too simple to describe this system. The KRR model was relatively successful, as it was able to predict the test set quite well based on the r2 scoring metric. The Lasso model showed surprisingly poor performance, given how well the KRR model performed. The r2 score of its prediction was relatively constant around 60%, showing little variation with hyperparameters that tended towards zero until producing errors. Theoretically,

the only difference between KRR and Lasso is the way of weights regularization. KRR uses L2 norm while Lasso uses L1 norm. L1 norm will force the small weights to zero, so the Lasso model will drop some features during optimization, which reduces the number of features furthermore. Given that the number of features might be insufficient in the multi-linear model, the reason why Lasso's performance was not satisfactory might be underfitting the data due to too few features.

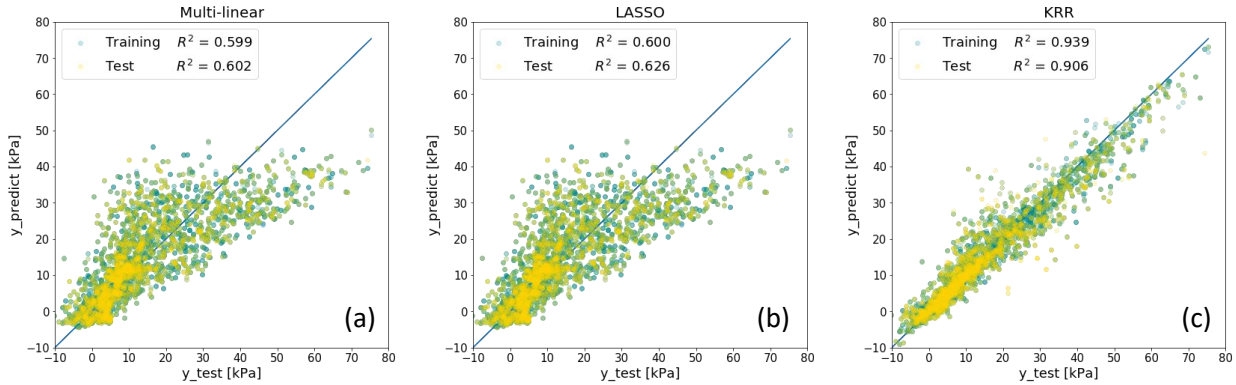

Figure 6. Comparison of different models. (a) Simple multi-linear regression; (b) LASSO, optimal α=0.001 σ=0.01; (c) Kernel ridge regression, optimal α=0.1 σ=0.5.

Given that the data is a time-series dataset, we also calculated autocorrelation using the time-series model, and we found that 2 prior points having significant partial autocorrelations, but no long-term seasonal variations were visible. This agrees with the original experimental setting, as when doing the experiments, the different states are randomly chosen thus they have no time dependent. Also. because we expect the system reaches equilibrium relatively quick, the time dependency decays quick enough for us to ignore. We see some scatter data points prior to 5 steps before also have high partial correlation, this probably is due to the coincidence from experimental design. Overall, we have illustrated that the time-series model is not necessary for our model, and in the following models, we will discard the anytime dependency by shuffling the data. Moreover, as a baseline model, the KRR model's success definitely bears continued attention and optimization as we progress.

MODEL IMPROVEMENT

Hyperparameter tuning

As a winning model in last section, the KRR model was chosen to be improved by tuning the hyperparameters. Wide ranges of hyperparameters and a secondary GridSearch were employed in the tuning step. The final, optimal hyperparameters are: alpha = 0.01 and sigma = 0.3, which increase the R score of the training set to 0.999 and R score of the test set to 0.987.

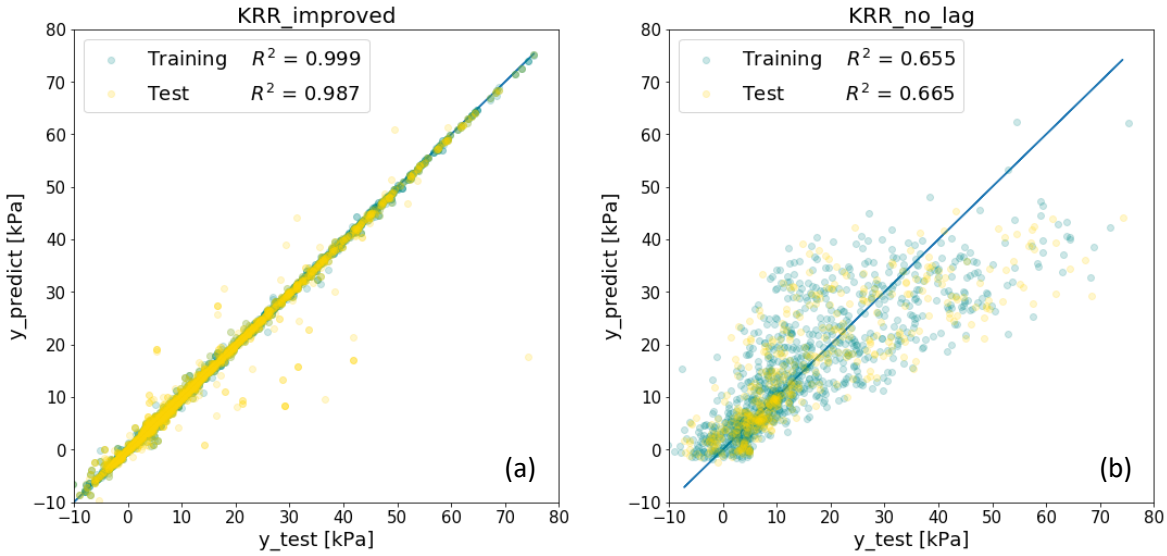

Figure 7. Comparison of original data and non-lag data. (a) Improved KRR model with original data after hyperparameter optimization, optimal α=0.01 σ=0.3; (b) KRR model with the no-lag data after hyperparameter optimization, optimal α=0.05 σ=5.

Non-lag data

As can be seen in Fig.7, we implemented a modified dataset named no-lag data. This data is created to solve the problems we observe in our original data, namely 1. there exists lag between pressure change and flow change during the experiment for some data points and 2. There exists lots of replicated data points.

Note that instead of being based on data analysis theory, we identified this issue based on chemical engineering knowledge. The data collected with lag is different than that collected at steady state; however, including replicated data points has no contribution towards a meaningful model. We dealt with this problem by removing any duplicate data points, opposed to factoring in a lag. After doing this, the number of data points decreased to roughly 1500.

From the regression result, we find that the R score gets worse in KRR model. This may be mainly due to the deletion of replicated data points that were being weighted more heavily and consequently fit more accurately.

CONCLUSION

Four regression models were implemented to predict downstream pressure, showing various levels of success. 1D regression failed as it was not able to accurately account for the dynamics of the system arising from varying two-phase flow regimes. Multi-linear regression failed to predict much of the data in higher pressure regions, demonstrating its inability to capture some of the system's non-linearities. Lasso regression failed in a similar fashion, failing to capture the system's trends in high-pressure regions despite attempts at hyperparameter tuning; however, expectations for the Lasso model were low, given the original data set contained only 8 features and dropping any would have lost a significant amount of unique information. The KRR model did succeed in predicting P6 at all regions of the system after tuning of the hyperparameters. Similar r^2 values were produced for training and testing datasets, demonstrating that the model was not overfitting the training dataset; however, the high r^2 values may have been skewed by nearly replicate data points in the system as produced by steady-state operation. Even KRR, the best predictive model for the entire dataset, showed significantly decreased performance on a 'no-lag' dataset without replicated points. Time-series modelling may hold more potential for modelling start-up operations more accurately; however, it was predominantly neglected for two reasons. First, the known physical features of the system are not expected to vary significantly with time, and second, the system reached steady state relatively quickly, removing much of the need for time dependence. Ultimately, we would recommend that separate, further data collection and modelling be implemented if accurate predictive modelling for start-up or shifting steady state process conditions is desired, but KRR modelling is quite successful at modelling the entire system, including steady state operations.

INDIVIDUAL CONTRIBUTIONS

	Lisa Reisenauer	Triton Wolfe	Joshua Randrup	Zhenzi Yu	Yuhan Yang
Project Definition	X	X	X		
Data Processing				X	X
Baseline Model	X		X	X	X
Improved Model	X	X	X	X	X
Final Draft	X	X	X	X	X
Presentation	X	X	X	X	X

BIBLIOGRAPHY

Curtis Ingstad Carlson, J. H., Michael Stanley DeCourcy, H. T., & Jamie Jerrick John Juliette, H. T. (2007, August 07). United States Patent No. US 7.253,307 B1.