Classification of Ultrasonic Flowmeter Health Diagnostics Using a Random Forests Approach

Ryan Christoper R. Dajay
Electronics and Communications Engineering Department
De La Salle University - Manila
Taft Avenue, Manila, Philippines
ryan_christoper_dajay@dlsu.edu.ph

***Abstract*—In this paper, the random forest classifier was used to model the health diagnostics of an ultrasound flowmeter. The dataset consisted of *181* samples with *43* different diagnostics features and *1* class output for the health diagnostic of the flow meter. After implementing k-folds cross-validation, average results showed low mean squared error (*MSE*), mean absolute error (*MAE*), and root mean squared error (*RMSE*) and with an average accuracy of *81.72%*. Due to the large number of features, analysis of relevant features was also done in the study. A potential area of future research is dimension reduction and feature engineering of the large dataset and exploring the results with using different optimization for ensemble models to further improve the accuracy.**

Keywords—random forests, regression trees, ensemble, classification, machine learning, ultrasonic flowmeter

I. INTRODUCTION

Decision tree is a simple and widely used classification technique. The application of decision trees ranges extensively from the fields of word recognition, medical problem diagnosis, variable selection, assessing the relative importance of variables, prediction and data manipulation [2][3]. The advantages of using decision tree classifier includes ease of interpretation, can be applied in classification and regression, and, are flexible and robust pattern recognition technique. However, decision tree classifiers are prone to overfitting. This produces models that are able to make accurate predictions with the current dataset, but are not able to provide reliable predictions for additional data.

Since decision trees are prone to overfitting, there are numerous ways to approach this problem such as pruning, and ensemble learning methods such as bootstrap aggregation and boosting. For this study, random forest classifier, an ensemble model based on multiple decision trees with bootstrap aggregation will be used. [12][13] [14]. Present applications of random forest are seen in biomedical and agricultural fields for regression and feature analysis. [15][16]

Ultrasonic flowmeter is commonly used for volumetric flow measurement for either liquid or gases. In essence, the flow meter measures the velocity of a fluid or gas with ultrasound to determine volume flow. There are two commonly methods seen in flow meters: time transit method and Doppler method [11]. For the time transit method, it works based on the propagation velocity of sound waves. Two sensors are used to transmit and receive signals simultaneously. At zero flow, there will be no transit delay between the sensors. For the Doppler method, it incorporates the Doppler shift that results from reflection. For this study, the time transit method was used. Due to the delicate process in determining the velocity of the fluids, it is important to continuously monitor the health diagnostics of the meter. This focus of this paper is to create a diagnostic system for ultrasound flow meter with a machine learning approach instead of a rule-based expert system approach.

II. CONCEPTUAL FRAMEWORK

A. Decision Tree Learning

Decision tree is a model of prediction and classification which incorporates supervised type of learning to determine the relationship of the input features and outputs [5][6]. The structure of the decision tree is a tree-like flowchart that is generated from top to bottom that can be seen in Fig.1 below. Due to how it is structured, decision trees provide easy visualization and representation of relative importance of the input features to its outputs. The fundamental node is called the root node, and at the end of the tree is called the leaf node [1]. From the fundamental node, the model will split out at decision nodes to ultimately end at a leaf node. In building a decision tree, one of the popular algorithms is called classification and regression tree (CART) algorithm [4]. The CART uses Gini index to determine the degree of impurity. Furthermore, the splitting point is the point of minimum Gini index. For this study, instead of using the Gini index, the mean squared error (MSE) will be used as the decision criteria to decide to split a node in two or more sub-nodes.

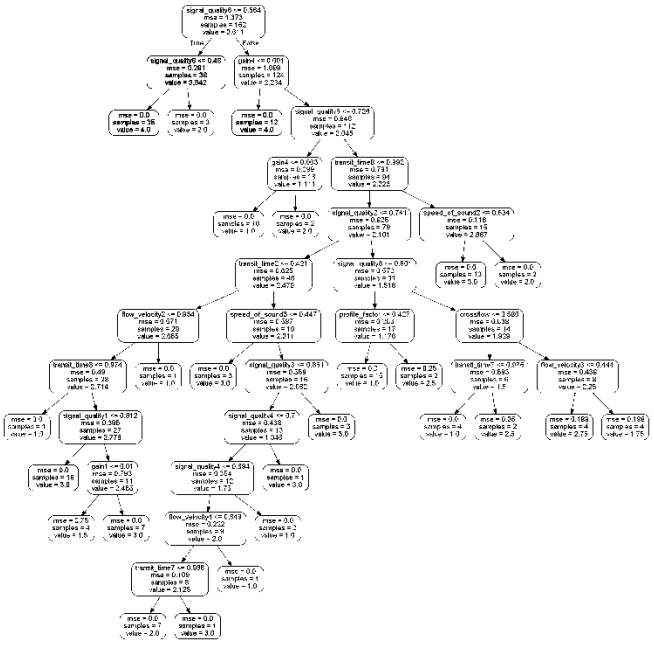

Fig. 1. Sample of a Decision Tree

In general, the size of the tree depends on the number of input features in the dataset. However, the main drawback of using decision trees is that they are prone to overfitting. Due to overfitting, pruning must be done in the decision tree model to improve the predictive accuracy of the model, while, at the same time, reducing the overfitting. In essence, pruning removes the section of the decision trees that has a low contribution to classify instances.

Many researchers had also shown that ensemble methods to boost accuracy for unstable learning algorithms such as artificial neural network (ANN) or decision trees; and, have shown that pruning is intended to make decision trees simpler and more comprehensible, but does not necessarily lead to improved generalization [7].

For this study, instead of using pruning techniques, models incorporating ensemble methods such as bagging and boosting will be used to improve the predictive accuracy.

B. *Random Forests*

The random forest model also incorporates supervised type of learning based on ensemble learning. In ensemble learning, multiple same or different types of algorithm are joined together to form a more powerful prediction model. In this case, multiple decision trees with each its own fundamental nodes, decision nodes, and leaf nodes are built.

For this study, random forest with bootstrap aggregation or bagging will be used for the dataset. Bootstrap aggregation is a machine learning ensemble algorithm that reduces overfitting and variance of the dataset, and, at the same time, improves the stability and accuracy of the model. In essence, the idea of bagging is to average the predictions of multiple independent models trained with different subset of the given dataset to obtain a model with a lower variance.

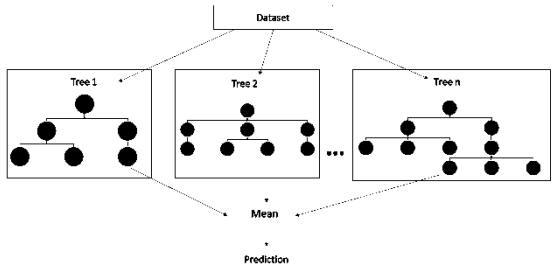

Fig. 2. Sample of Random Forest structure and with ensemble method algorithm

The advantages of using random forest includes excellent performance in terms of classification and predictive tasks even with handling large numbers of variables with small sets of observations with high accuracy [8][9].

III. METHODOLOGY

A. *Data Acquisition and Pre-processing*

A total of 181 samples, with 43 features and 1 class, describing the diagnostics of ultrasonic flowmeter were acquired from the online repository of Center for Machine Learning and Intelligent Systems at the University of California, Irvine [10]. The dataset includes the profile factor, symmetry, crossflow, flow velocities in each of the four paths, speed of sound in each of the four paths, signal strength in each of the four paths, signal quality in each of the four paths, gain at both ends in each of the four paths, transit time in each of the four paths as the input features and, lastly the health state of the meter for the given input features was also given. The health state of the meter given the diagnostics is described by Table I.

The dataset was normalized using *MinMaxScaler* function. It works by getting first the range of the original maximum and minimum value of a feature. And, for every feature, it is then subtracted by the minimum value and divided by the range. The advantage of using this pre-processing technique: it preserves the shape of the original distribution.

TABLE I. CLASSIFICATION OF ULTRASONIC FLOWMETER DIAGNOSTICS

Class	Specification
Healthy	Ultrasonic flowmeter is operating under normal conditions.
Gas Injection	Injection of unwanted gas that affects the velocity performance of the flowmeter
Installation Effects	Increase of uncertainty due to the “non-ideal” on-site installation
Waxing	Formation of wax deposits in the flowmeter

B. *Tools and Software*

Python was the choice of software used for classifying the diagnostics of the ultrasonic flowmeter dataset. The *sckit-learn* library contains several tools for machine learning analysis. Among these tools, *RandomForestRegressor* was used to model the dataset. The accuracy metrics, and tree diagram of the model was also implemented using the same library.

C. *Program Framework*

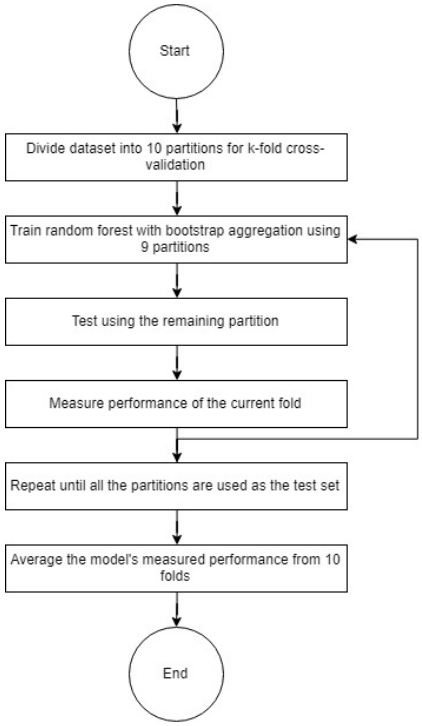

Fig. 3. Program framework of the study

To guarantee the performance of our model, cross-validation method was implemented during training and evaluation for our model. The dataset contains 181 samples that was divided into 10 partitions to perform k-fold cross validation with 10 folds. With each fold, 1 unique partition is used as the test set and the remaining 9 partitions were used for training. The average accuracy and error metrics was then acquired after 10 folds.

D. Model Hyperparameter Tuning

TABLE II. RELEVANT HYPERPARAMETERS OF RANDOM FOREST

Class	Specification
N_estimators	represents the number of trees in the forest
max_depth	represents the depth of each tree in the forest
min_samples_split	represents the minimum number of samples required to split an internal node
min_samples_leaf	represents the minimum number of samples required to be at a leaf node
max_features	represents the number of features to consider when looking for the best split.

For the Random Forest, the important hyper-parameters of the model are listed in Table II. The hyper-parameters of the model are tuned using the dataset to improve the accuracy of the model.

IV. ANALYSIS OF RESULTS

A. Feature Importance

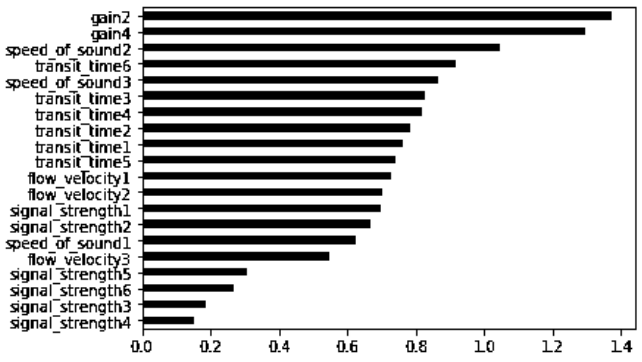

Fig. 4. Features with the least relevance

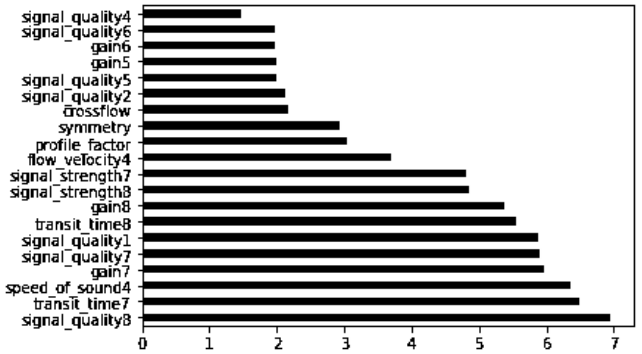

Fig. 5. Feature with the most relevance

Fig. 4 and Fig. 5 present the importance of features in the given dataset. It is the measure of its relevance toward the output variable. From the figures, it can be seen that the least relevant input features shows relevance of ranges below *1.4%*, while the other features with the most relevance shows to have a range of relevance greater than *1.4%.*

B. Accuracy and error metrics

The performance metrics of the random forest classifier is listed in Table III.

TABLE III. RANDOM FOREST CLASSIFIER PERFORMANCE METRICS

Metric	Value
Mean Accuracy	81.72%
Mean Absolute Error (MAE)	0.0994
Mean Squared Error (MSE)	0.0269
Root Mean Squared Error (RMSE)	0.1574

Equations (1) to (3) below presents the mean absolute error (MAE), mean squared error (MSE), and root mean squared error (MSE) respectively. These will be used to determine the overall performance and accuracy of our trained model.

Mean absolute error (MAE) measures the average magnitude of the errors in a set of predictions without the consideration of the direction of their error. It is the most intuitive of the given metrics since it is just the absolute difference of the data and the trained model's prediction. Due to this, it does not indicate if the model is underperforming or over performing.

Mean Squared Error (MSE) is similar to the mean absolute error, but the error is squared before summation of all errors. The presence of the squared term makes the MSE better in analyzing the performance of the model with respect to the outliers of the dataset.

Lastly, the root mean squared error (RMSE) is similar with the MSE and is analogous with the standard deviation of the trained model.

$$MAE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \quad (1)$$

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \quad (2)$$

$$RMSE(y,\hat{y}) = \sqrt{\left(\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2\right)} \quad (3)$$

C. Tree Plot

Fig. 6 to Fig. 8 shows tree sample plots for the random forest model. As discussed above, random forest model is the prediction of multiple decision trees that are then averaged to keep the variance and overfitting low. It can also be seen that the mean squared error (MSE) reduced as you go down the tree. This is expected, as the splitting criteria for building a tree requires to have the least MSE.

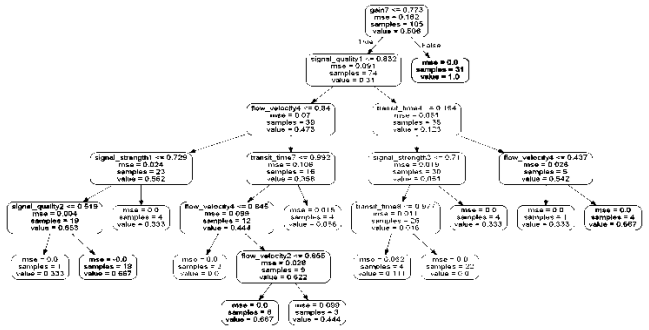

Fig. 6. Tree plot sample I in Random Forest

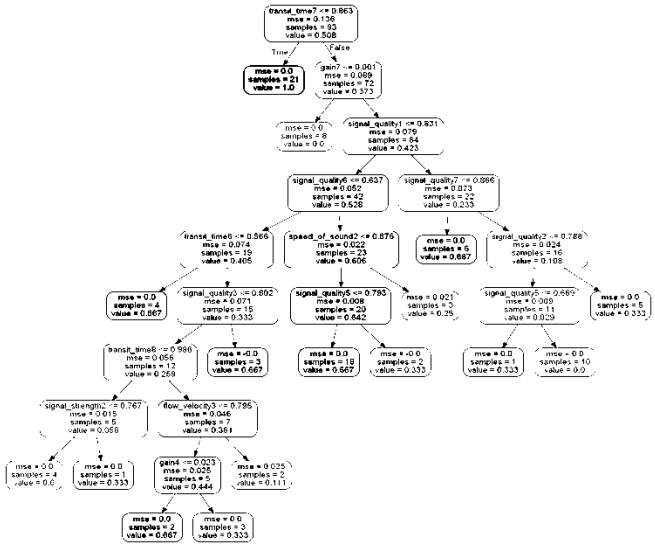

Fig. 7. Tree plot sample II in Random Forest

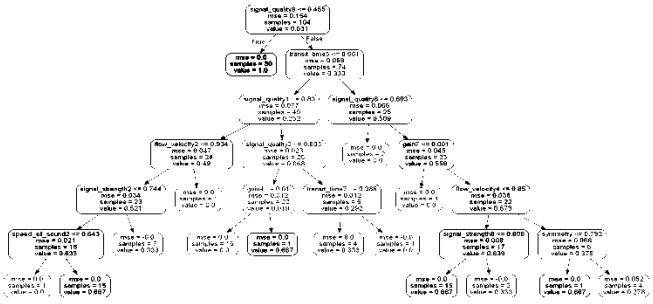

Fig. 8. Tree plot sample III in Random Forest

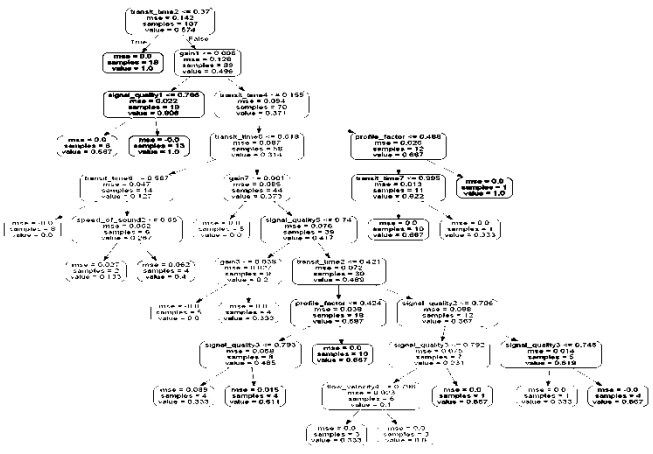

Fig. 9. Tree plot sample IV in Random Forest

V. CONCLUSIONS AND RECOMMENDATIONS

This paper presented a machine learning approach for the classification of ultrasonic flowmeter diagnostics instead of the traditional approach which is a rule-based expert system. Acceptable performance metrics have been attained using the random forest model. Mean accuracy of 81.72% have been recorded with low mean absolute error, mean squared error and root mean squared error. It shows that classification using random forest is an effective approach in classifying the health state of the ultrasonic flowmeter based on its diagnostics. Additionally, the relevant features in the diagnostics have also been explored in the paper that can be used as basis for future feature engineering to improve the accuracy of the model's prediction.

Future work may also involve comparing the random forest model with other machine learning techniques such as artificial neural network (ANN), stochastic gradient descent (SGD) or support vector machines (SVM). Alongside with other learning models, different optimization and ensemble methods can be explored to boost the accuracy of the model.

REFERENCES

[1] Chen, J. R., Lin, Y. H., & Leu, Y. G. (2018). Predictive model based on decision tree combined multiple regressions. *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, (1), 1855–1858. https://doi.org/10.1109/FSKD.2017.8393049

[2] E. Kaya, B. Barutçu, and S. Menteú, "A method based on the van der Hoven spectrum for performance evaluation in prediction of wind speed," Turk. J. Earth Science, vol. 22, pp. 1–9, 2013.

[3] [10] S. Kannan , and S. Ghosh, "Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output," Springer-Verlag, July .

[4] L. Breiman, and J. H. Friedman, R. Olshen, and C. J. Stone. Classification and Regression Trees. Chapman & Hall, New York, 1984.

[5] E. G. Petre, "A decision tree for weather prediction", Buletinul, vol. LXI, no. 1, pp. 77-82, 2009.

[6] P.Hemalatha, "Implementation of data mining techniques for weather report guidance for ships using global positioning system,"International Journal Of Computational Engineering Research, vol. 3, no. 3 , March 2013.

[7] Famili, A. F., Kok, J. N., Peña, J. M., Siebes, A., Feelders, A., Gunetti, D., … Ruffo, G. (2005). *Advances in Intelligent Data Analysis VI. 3646*(August 2001), 1 a33-144–144. https://doi.org/10.1007/11552253

[8] G. Ulrike, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," The American Statistician, vol. 63, no. 4, January 2009, pp. 308-319.

[9] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017). The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data. *2017 14th International Conference on Engineering of Modern Electric Systems, EMES 2017*, 1–4. https://doi.org/10.1109/EMES.2017.7980368

[10] K. S. Gyamfi, J. Brusey, A. Hunt, E. Gaura , â€œLinear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics,â€ Expert Systems with Applications (IF: 3.928), September 2017

[11] Vermeulen, M. J. M., Drenthen, J. G., & Den Hollander, H. (2012). *Understanding diagnostic and expert systems in ultrasonic flow meters*. 1–32. Retrieved from https://pdfs.semanticscholar.org/1b4a/d8d8cc568cfc9fa5fa68f79689b4da3abc52.pdf

[12] Liaw, A., Wiener, & Mathe. (2004). Classification and regression by randomForest. *Journal of Dental Research*, *83*(5), 434–438. https://doi.org/10.1177/154405910408300516

[13] Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. *Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017*, 65–68. https://doi.org/10.1109/WCCCT.2016.25

[14] Doan, A. A. Q. (n.d.). *Intro To Random Forest*.

[15] Bashar, S. S., Miah, M. S., Karim, A. H. M. Z., Al Mahmud, M. A., & Hasan, Z. (2019). A Machine Learning Approach for Heart Rate Estimation from PPG Signal using Random Forest Regression Algorithm. *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, 1–5. https://doi.org/10.1109/ECACE.2019.8679356

[16] Bravo, S., & Moreno, A. H. (2019). A Random Forest Approach for Predicting the Microwave Drying Process of Amaranth Seeds. *2019 IEEE 2nd International Conference on Information and Computer Technologies, ICICT 2019*, 25–29. https://doi.org/10.1109/INFOCT.2019.8711122