Machine learning algorithms for classification of boiler faults using a simulated dataset

Rony Shohet¹, Mohamed S Kandil¹ and J J McArthur¹,*

¹ Dept. Architectural Science, Ryerson University, 350 Victoria St., Toronto, ON, Canada, M5B 2K3
* jjmcarthur@ryerson.ca

Abstract. Building performance has been shown to degrade significantly after commissioning, resulting in increased energy consumption and associated greenhouse gas emissions. Continuous Commissioning using existing sensor networks and IoT devices has the potential to minimize this waste by continually identifying system degradation and revising control strategies to adapt to real building performance. Due to its significant contribution to GHG emissions, building heating, particularly gas boiler systems are critical systems for detecting decreased performance. A review of boiler performance studies has been used to develop a set of common faults and degraded performance conditions, and these have been integrated into a MATLAB Simulink emulator to create a labelled dataset with approximately 27,500 cases for training and testing boiler fault classification models. Classification algorithms such as K-nearest neighbour, Decision tree, Random Forest and Naïve Bayes have been tested and the results show that decision tree methods gave the best prediction (97.8% accuracy) followed by Random forest (95.0%) and KNN for K = 3 (88.1%). Naïve Bayesian and KNN for K = 9 classification both gave poor results.

1. Introduction

HVAC systems throughout a building’s life cycle may often lead to poor performance due to faulty equipment. Amongst all end-uses in buildings, heating, ventilation, and air conditioning (HVAC) accounts for 40% of building energy consumption [1]. Common faults in HVAC equipment includes process parameter changes, disturbance parameter changes, actuator problems, and sensor problems [2]. These faults may accumulate over time, often undiagnosed, resulting in decreased performance and increased energy consumption and costs. Fortunately, fault detection and diagnosis (FDD) technology can leverage this understanding of poorly operating equipment to improve performance. The goal of FDD includes improved indoor environmental quality, reducing unscheduled equipment down time and maintenance costs, and increased equipment life [2]. However, accurate FDD requires detailed knowledge of how faults affect the performance of the system either with recorded sensor data or through fault modelling [1]. Li O’Neill note that the development of fault data using simulation is extremely valuable as it permits the modelling and algorithm training for complex fault scenarios (multiple concurrent faults) and is a way to inexpensively generate the bulk data necessary for algorithm development and testing [1]. Further, this approach permits data on rare or dangerous fault conditions to be generated without risk to the building or its occupants. To address the gap noted in [1] regarding a dearth of simulation-based studies, this paper presents a study of the operation of an HVAC component through simulation and modelling. Using MATLAB/Simulink and Simscape, the subcomponents within a boiler can be implemented and designed as such to simulate nominal and faulty performance. It was found that such accumulation of faults throughout an entire HVAC system can contribute to an additional 40% in energy consumption [1].

A comprehensive literature review of fault modelling in HVAC systems is presented in [1]. Fault simulation can be categorized as three groups: white-box (physics, first principles), black-box (data driven, machine learning, empirical) and grey-box (hybrid, semi-empirical) [2]. White box models
utilize concepts of physical and chemical laws, such as mass, momentum, and energy balance to develop relationships between inputs and outputs [3]. Simulation software exists that can model individual HVAC components, ranging from primary components (e.g. boiler, heat pumps, and chillers) to secondary components (e.g. dampers and air handling units), or entire HVAC systems [4]. The presented research utilizes concepts of white box models to develop the emulator of a boiler. Black-box models are data driven, relying on empirical solutions, especially machine learning concepts. Regression, fuzzy logic, frequency domain, and other similar approaches are most commonly used for HVAC system modelling [4]. The third broad model type is a ‘grey box’ model, which combines both physics-based and machine learning approaches.

This paper contributes to the development of improved fault detection in two ways: first, it presents a validated physics-based boiler emulator model that can be used to generate simulated data for rare fault conditions, and second, it explores machine learning models for fault detection using points typically monitored by Building Automation Systems (BAS). The dataset associated with this data is also provided as a supplemental file to complement field-collected data and support future research.

2. Methodology

This paper presents a combined approach, whereby the physical system is simulated within MATLAB and modified from a nominal case based on the Simscape heating system model [5], a toolbox within the Simulink library capable of modelling heating systems and validated using manufacturer data. To simulate potential fault conditions, model parameters are varied, either individually or in combination, and the resulting input and output data is labelled with the fault name. This results in a set of datasets, which can be filtered to mimic the point outputs from a BAS and used to train machine learning models to classify sets of BAS points to detect or predict fault conditions. The resultant dataset can be used to supplement logged data where certain faults have not yet occurred and thus the real-world data is unavailable. There were three fundamental steps in this research: (1) development and validation of a boiler emulator capable of simulating normal operation and operation under key fault conditions; (2) simulation of fault condition dataset; and (3) creation of a testing and training dataset for a machine learning model to identify fault conditions based on standard building monitoring system data points.

2.1 Emulator development and validation

The components modelled within the emulator include the combustion chamber and the gas/water heat exchanger. First, a component of a combustion chamber was taken from a published Simscape model [5]. This combustion chamber represents the area where the combustion fuel, natural gas for this model, undergoes combustion with air. Following the design and selection of inputs, a heat exchanger component was connected to the combustion chamber. The heat exchanger represented the latent heat transfer between combustion gasses and passing water. The water was passed through the system at a constant mass flow rate through the Water Mass Flow Rate Source component. The heated water was then directed into a Reservoir component, which acts as the Supply Water source for a building. The block diagram the cumulative system is shown in figure 1.

The constant parameters that have been modified are tabulated for each component in table 1, to represent the Viessmann Vitorond 200 Gas Fired Boiler VD2 Series 380. This model was validated by replicating the ANSI/AHRI Standard 1500 – Performance Rating of Commercial Space Heating Boilers [10] test conditions and comparing the predicted outputs with the published manufacturer data [8]. Non-constant values such as ambient temperatures and boiler loop flowrates acted as the changing parameters to generate a diverse dataset. Table 2 shows the validation results achieved. The expected (real equipment) output for CO2 is based on natural gas combustion while achieved (emulator) output is based on methane, the primary – but not sole – component of natural gas. This explains the discrepancy in validation results. In addition, the underlying assumptions of this model, including steady state operation, adiabatic boiler enclosure, and sensible-only heat transfer.
Figure 1. Boiler emulator developed in Simscape.

Table 1. Summary of constant parameters.

| Component                        | Parameter                          | Value and Source |
|----------------------------------|------------------------------------|------------------|
| Boiler                           | Hydrocarbon lower heating value    | 50 MJ/kg [6]     |
|                                  | Fuel specific heat at constant pressure | 2191 J/kg/K [6] |
|                                  | Dry air specific heat at constant pressure | 1005 J/kg/K [7] |
| Gas/Water Heat Exchanger         | Flow arrangement                   | Shell and Tube [8] |
|                                  | Number of shell passes             | 2 [8]            |
|                                  | Wall thermal resistance            | 0 K/W [9]        |
|                                  | Hydraulic diameter for pressure loss | 101.6 mm [8]   |
|                                  | Thermal diameter for pressure loss | 0.275 m³ [8]    |
|                                  | Thermal Liquid - Heat transfer surface area | 10.1 m³ [8] |
|                                  | Controlled Fluid - Heat transfer surface area | 18.2 m³ [8] |
|                                  | Thermal Liquid - Initial temperature | 333 K [8]      |
| Water Mass Flow Rate Source      | Cross-sectional area at ports A and B | 0.1 m² [8]   |
| Reservoir – Return Water         | Reservoir temperature              | 333 K [8]        |
| PS Constant – Temperature Fuel   | Constant (Temperature)             | 300 K [8]        |
| PS Constant – Humidity Ratio     | Constant (Humidity Ratio)          | 0.001 kg/kg [5]  |

Table 2. Validation results.

| Validation Parameter           | Component                  | Expected Output | Achieved Output |
|--------------------------------|----------------------------|-----------------|-----------------|
| Water Return                   | Reservoir - Return         | 333 K [8]       | 333 K           |
| Water Supply                   | Reservoir - Supply         | 348 K [8]       | 348 K           |
| Boiler Heat Output             | Combustion Chamber         | 387 kW [8]      | 405 kW          |
| Combustion Products            | Combustion Chamber         | 10% CO₂ [10]    | 9.5% CO₂        |

2.2 Fault condition simulation dataset development

Once the boiler emulator had been validated, modifications to input conditions were incorporated to simulate a range of common boiler faults and thus generate a dataset intended to complement logged data from a BAS. The generated datasets were split 80% training / 20% testing using a random seed, which was consistent for all algorithms tested. While a variety of factors were modified within the emulator, summarized in table 3, to reflect the physical causes of each fault, only those data points
visible to the BAS, namely the water flow rate, entering and leaving water temperatures, outdoor air and fuel temperatures, and gas consumption rate, were output to the dataset and were labelled with the associate fault. Iterations were performed changing the gas fuel rate from 1 kg/s to 4 kg/s, water mass flow rate from 3 kg/s to 12.5 kg/s and combustion air temperature from 283 K to 303 K. A constant return temperature of 333K was used for all runs and thus omitted from the dataset. A total of 27,281 simulations were run to generate a robust dataset [11] for model training.

| Fault (Label)         | Component                        | Variable         | Nominal Value | Tested Range   |
|-----------------------|----------------------------------|------------------|---------------|----------------|
| Excess air (X)        | Boiler combustion chamber         | Air flow rate    | 0             | 5% - 50%       |
| Gas-side fouling (F)  | Gas-Water heat exchanger          | Fouling factor (%)| 0             | F = 0.01 - 0.46|
| Water-side Scaling (S)| Gas-Water heat exchanger          | Scaling factor (%)| 0             | S = 0.01 - 0.46|

2.3 Classification using machine learning
Consistent with best practices for machine learning research, a selection of common classification algorithms were tested for their ability to distinguish between normal operation and each fault, namely K-nearest neighbour (KNN), Naïve Bayes (NB), Decision Trees (DT), and Random Forest (RF) using 500 trees. The algorithms were programmed in R [12], which was used for model training, testing, evaluation, and result visualization. To allow the progression of a fault condition to be detected over time, for example fouling increasing from 1% to 6%, the full set of conditions (31 classes) was used for the initial classification. In addition, a broader classification by type of fault {Excess air, Fouling, Scaling} versus Nominal operation was also tested. Finally, feature selection was used to improve results.

3. Results and Discussion
Figure 2 shows the results for the categorical classification. The prediction accuracy of KNN improved consistently, increasing to 91.0% for k=3 and 65.3% for k=9. Since most of the misclassification occurred within adjacent faults, this improvement in accuracy is a result of combining each individual fault and their misclassifications as one broader class. The Naïve Bayes algorithm also improved with the decreased number of classes, from 5.8% to 46.7% testing accuracy. DT maintained a high level of accuracy 97.2%, while RF fell to 73.6% and there was noticeable confounding was a result of fouling being misclassified as scaling.

The full condition classification results are shown in figure 3. Of the algorithms, DT had the highest accuracy (97.8%) followed by RF (95.0%) and then KNN with k=3 (88.1%). A thorough analysis of the results showed that RF consistently amplified the misclassifications occurring to a lesser degree in DT. For example, RF misclassified X=0.1 as S=0.01 for 34.0% of occurrences compared with 2.2% for DT. In addition, when feature selection was implemented, RF misclassified adjacent excess air faults, as well as misclassifying between faults and scaling. This may be a result of removing gas mass flow rate, as it would have provided insight capable of distinguishing between similar fault outputs. Naïve Bayes and
KNN with larger number of neighbours (k>5), performed poorly for all feature sets tested, likely due to the **curse of dimensionality** associated with such a large number of classes. It is noteworthy that of all the algorithms tested, the KNN model showed the most significant performance improvement with feature selection, with the k=3 model increasing from 4.3% to 88.1% when fuel rate was removed as an input. Beyond k=3, the accuracy for this model remained consistently poor regardless of input variables. Conversely, the random forest model suffered, decreasing from 95.0% to 74.2% when the fuel flow rate was omitted. The remaining algorithms showed no such sensitivity to feature selection.

**Figure 3.** 31-Class confusion matrices for best feature set algorithms tested: KNN with k=3 (top left), NB (top right), DT (bottom left), and RF with 500 trees (bottom right).

Despite the large number of classes, the condition prediction was deemed to be successful for fault detection, particularly the DT model with 97.8% accuracy. This granularity in prediction is important because it permits a more precise diagnosis of the specific fault occurring within the boiler. Further, if left unresolved, it is possible to track the extent of the fault in time, and thus build future models permitting a mean time to failure estimate to be developed. Together, these algorithms will permit an intelligent boiler monitoring system to be developed and integrated into the building automation system, thus providing an additional depth of insight into boiler fault progress, allowing for improved maintenance schedules and permitting the optimization of operational costs.
4. Conclusions
This study has determined that it is possible to classify faults across a large number of conditions with high accuracy based only on observed BAS data points. While presenting promising results, there are several limitations of this research as-presented. First, the boiler validation and testing was based on a single boiler model and future research should repeat the validation testing for other boiler models and create similar datasets for those boilers. Second, the classification is only performed for individual faults not combined/hybrid faults. Third, this research presents only simulated results, and should be extended in the future to include field-collected data. To address the first two limitations, future work will clone this emulator to develop datasets for other boilers and replicate this study across boiler types (condensing and non-condensing) and sizes. Multiple concurrent faults will be simulated to permit more complex investigations to be undertaken. To address the third limitation, the authors are obtaining real data from in-situ boilers on campus and this data will be used to both enhance the dataset as well as further refine and validate the fault detection models. Additional studies are investigating the impact of signal noise on prediction accuracy and identify signal processing techniques to increase the robustness of the model for real-world applications.

References
[1] Li Y and O’Neill Z 2018 A critical review of fault modeling of HVAC systems in buildings. Build. Simul. 11 953-75
[2] Lan L and Chen Y 2007 Application of modeling and simulation in fault detection and diagnosis of HVAC systems Build. Simul. 1299-1306
[3] Homod Z R 2013 Review on the HVAC system modeling types and the shortcomings of their application Journal of Energy 2013
[4] Afram A and Janabi-Sharifi F 2014 Review of modeling methods for HVAC systems. Appl. Therm. Eng. 67 507-19.
[5] MathWorks 2019 House Heating System. Accessed January 1, 2019. https://www.mathworks.com/help/physmod/hydro/examples/house-heating-system.html?searchHighlight=heating&s_tid=doc_srchtitle
[6] Turns S 1996 An Introduction to Combustion (New York: McGraw-Hill)
[7] Satoh M. Atmospheric circulation dynamics and general circulation models. Springer Science & Business Media; 2013 Jul 4.
[8] Viessmann 2018 Viessmann Vitorond 200 Technical Data Manual October. https://www.viessmann.ca/content/dam/vi-brands/CA/pdfs/commercial/vitorond_200-lg_tdm.pdf/_jcr_content/renditions/original.media_file.download_attachment.file/vitorond_200-lg_tdm.pdf
[9] Shah R and Sekulić D 2003 Fundamentals of Heat Exchanger Design (Hoboken: John Wiley & Sons)
[10] ANSI/AHRI 2014 2015 Standard 1500 for Performance Rating of Commercial Space Heating Boilers Standard, Arlington: Air Conditioning, Heating, and Refrigeration Institute
[11] Shohet R, Kandil M and McArthur J 2019 Simulated boiler fault data Toronto: IEEE Dataport. Available online: doi: https://doi.org/10.21227/ye8z-z608.
[12] CRAN The Comprehensive R Archive Network. Accessed 03 29, 2019. https://cran.r-project.org