Explainable Classification of Nuclear Facility Operational State Using Node and Region Importance for Sensor Networks

Jake Tibbetts

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-95
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-95.html

May 14, 2021

Explainable Classification of Nuclear Facility Operational State Using Node and Region Importance for Sensor Networks

by

Jake Tibbetts

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Stuart Russell, Chair
Dr. Bethany Goldblum, Co-chair

Spring 2021

The thesis of Jake Tibbetts, titled Explainable Classification of Nuclear Facility Operational State Using Node and Region Importance for Sensor Networks, is approved:

Chair Stuart Russell Date 5/10/21

Co-chair Bethany Goldblum Date 5/11/21

University of California, Berkeley

Explainable Classification of Nuclear Facility Operational State Using Node and Region Importance for Sensor Networks

Abstract

Explainable Classification of Nuclear Facility Operational State Using Node and Region Importance for Sensor Networks

by

Jake Tibbetts

Master of Science in Computer Science

University of California, Berkeley

Professor Stuart Russell, Chair

Dr. Bethany Goldblum, Co-chair

Distributed multisensor networks record multiple data streams that can be used as inputs to a machine learning model designed to classify proliferation-relevant operations at nuclear reactors. This work proposes methods to assess the importance of each node (a single multisensor) and region (a group of collocated multisensors) to model accuracy. This, in turn, provides insight into model explainability, a critical requirement of data-driven applications in nuclear security. To determine the importance of the various nodes and regions for a given classification problem, traditional wrapper methods for feature importance were extended to nodes and regions in a multisensor network. On a dataset collected at the High Flux Isotope Reactor at Oak Ridge National Laboratory by a network of Merlyn multisensor platforms, these methods were used to identify high value and confounding nodes and regions for classifying nuclear reactor operational state. Specifically, the nodes near the facility's cooling tower were identified as high value sources. When applied in conjunction with black-box classifiers such as neural networks, node and region importance can provide insight into an otherwise opaque classification model in the nuclear security domain.

Contents

List of Figures

List of Tables

Acknowledgments

I thank Dr. Bethany Goldblum for her unwavering commitment to and support of students like myself. I thank Dr. Stuart Russell for his excellent advice and mentorship throughout my time as a graduate student. I thank Dr. Chris Stewart and Jon Whetzel for their support during this project. I thank the other members of the Complexity Group - Arman, Milena, Araav, Samira, Rob, James, and Marcus - for making it a true pleasure to work on this project during such a hard year for all of us. I thank the MINOS venture and the Oak Ridge National Laboratory Staff at the High Flux Isotope Reactor, with a special thanks to Jared Johnson, Will Ray, Randall Wetherington, and Michael Willis. I would not have been able to complete this work without the help and support of all these wonderful people who have been a part of my journey over the last year.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Berkeley National Laboratory under Contract DE-AC02-05CH11231 and by the University of California, Berkeley under Award DE-NA0003180. The project was funded by the U.S. Department of Energy, National Nuclear Security Administration, Office of Defense Nuclear Nonproliferation Research and Development (DNN R&D).

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Chapter 1

Introduction

Nuclear facility monitoring has been a critical aspect of the modern international nuclear nonproliferation regime since the Treaty on Non-Proliferation of Nuclear Weapons came into force in 1970 [36]. Recent advances over the last decade in sensor technology and data science have created new opportunities to apply machine learning techniques to the problem of real-time nuclear facility monitoring for the purposes of nuclear security [12, 33, 11]. Since nuclear security is a high stakes domain, data-driven applications in this area must be both accurate and explainable so that analysts, policymakers, and decision makers trust the validity of the models [20, 4].

As sensor technology continues to improve, nuclear facilities are being monitored using networks of geographically distributed multisensors, also called nodes for shorthand, to record various data streams measuring physical phenomena. Machine learning techniques could be applied to these data streams to create models of proliferation-relevant signatures. Due to the high stakes nature of nuclear security, it is critical to explain in a human-understandable way how these models translate data streams into the desired proliferation-relevant signatures. This provides end-users, such as analysts and policymakers, the ability to understand these models' strengths and limitations which engenders trust in and increases the usability of these models in human decision-making processes [20]. One important aspect of model explainability is an understanding of which nodes and regions, where a region is a group of nearby collocated nodes, in a multisensor array are the most important for and have the largest impact on making accurate predictions.

This question of node and region importance is broadly applicable to any predictive task where sensor networks are being used to record data used as inputs to data-driven models. For these predictive tasks, node and region importance can be used to eliminate noisy features that reduce model performance which is a critical task for any machine learning application. Node and region importance can also be combined with knowledge about the specific problem context to make inferential hypotheses about the data and the classification models. For example, if a node in an area associated with a certain sensing spectrum is identified as having a strong positive impact on model accuracy and there is a domain-specific connection between that sensing spectrum and the predicted label, one could hypothesize that there

is a causal relationship between the label and the sensing spectrum which could be further exploited in future modeling efforts.

The goals of this work are to: (1) create machine learning models trained on sensor network data predicting nuclear reactor operational state, (2) apply node and region importance methods to these models, and (3) use the results of node and region importance analysis in conjunction with knowledge about the nuclear facility to improve model performance and make inferential hypotheses about the nuclear facility and the classification models. These goals were accomplished through the creation of a hidden Markov model (HMM) and a feed-forward neural network[1] predicting nuclear reactor operational state trained on data collected by a network of 12 geographically distributed Merlyn multisensor platforms deployed at the High Flux Isotope Reactor at Oak Ridge National Laboratory. These models were then analyzed with feature importance and selection wrapper methods extended to nodes and regions to assign importance scores to nodes and regions for each model. These importance scores were then leveraged to improve model performance further through feature selection and to make inferential hypotheses about the nuclear facility and the classification models providing insight into model explainability.

The accuracy of the HMM was increased from 0.583 to 0.839 through feature selection informed by node and region importance. Similarly the accuracy of the feed-forward neural network was increased from 0.811 ± 0.005 to 0.884 ± 0.004 through feature selection where the error bars are representative of a 95% confidence interval produced by training and evaluating a feed-forward neural network for 50 trials with different randomized initial weights and averaging the results. Additionally, node and region importance found evidence for a causal relationship between reactor operational state and the cooling tower at the facility. Node and region importance also found that the HMM was potentially sensitive to noise in the data produced by high foot and vehicle traffic which reduced model performance. Similarly, node and region importance found that the feed-forward neural network was sensitive to a sensor outage caused by a sensor component failure that reduced model performance. Node and region importance both informed the improvement of model performance through feature selection and provided insight into the resulting models and problem context improving overall explainability.

[1]Random forests were also examined. However the results were strictly worse than the feed-forward neural network performance and therefore were not reported.

Chapter 2

Dataset and Models

An HMM and a feed-forward neural network were trained on data collected by a network of 12 Merlyn multisensor platforms to predict binary nuclear reactor power state (off/on) at the High Flux Isotope Reactor at Oak Ridge National Laboratory.

2.1 High Flux Isotope Reactor

The High Flux Isotope Reactor (HFIR) [9] at Oak Ridge National Laboratory is an 85 MW research reactor used to study a variety of research questions in nuclear physics and engineering. Twelve multisensors were deployed at various locations (Fig. 2.1) around the facility which have been collecting data since April 2019. This work examines data collected over a time period of slightly less than 40 weeks. This 40-week period covers approximately six reactor power cycles which consist of a start-up, power generation at steady state for a period of time spanning approximately one to three weeks, and a shutdown.

There are a few relevant points of interest at the facility visible in Fig. 2.1. The main reactor building is at the center of the facility between Nodes 5, 6, and 8. The reactor's cooling tower is to the immediate right of Node 9. The target processing facility (the Radiochemical Engineering Development Center [10]) is between Nodes 4, 5, and 10, where targets placed inside of the reactor are prepared and disposed of after use. There are some liquid storage tanks immediately above Node 8. The main entrance to the facility is along the road where Nodes 1, 2, and 12 are deployed.

2.2 Merlyn Multisensor Platform

The Merlyn multisensor platform was designed by Special Technologies Laboratory, an organization within the Nevada National Security Site complex of facilities. A mockup of the Merlyn is shown in Fig. 2.2 and a picture of a deployed Merlyn at HFIR is shown in Fig. 2.3.

The platform was built with a BeagleBone Black mainboard [8], an ATmega328P-based Arduino UNO breakout board [1], a ROHM SensorShield EVK-003 sensor package [32], and

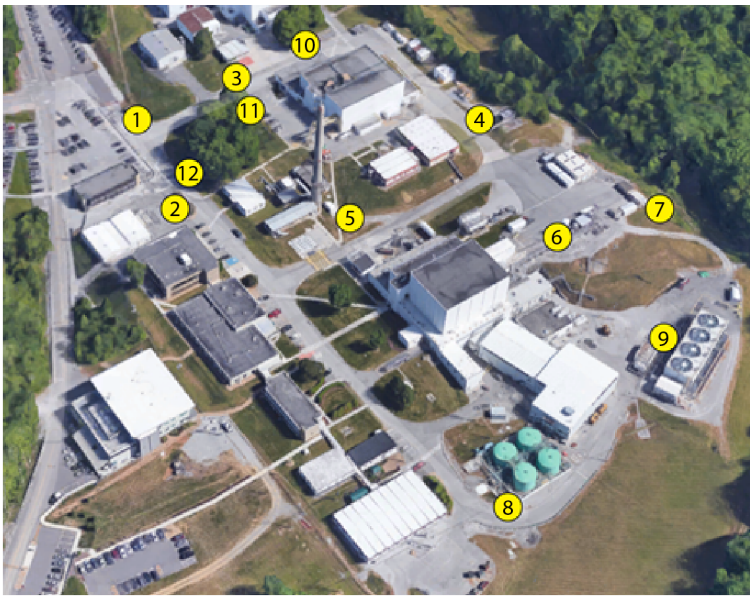

Figure 2.1: Overhead View of the High Flux Isotope Reactor With Labeled Node Locations

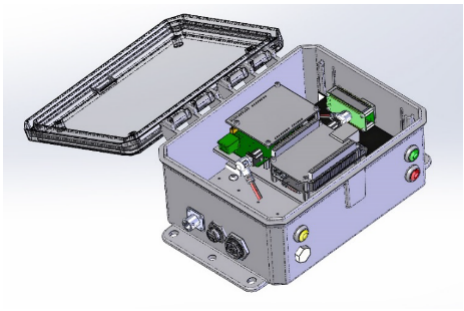

Figure 2.2: CAD mockup of a Merlyn inside a weather enclosure

Figure 2.3: Merlyn Deployed at HFIR

supporting hardware related to power and data distribution. The sensors in each Merlyn used in this work are listed in Table 2.1. Each sensor samples at a rate of 16 Hz.

2.3 Data Products

To create the data products used for modeling, a series of preprocessing transformations were applied to the raw data streams. First, each data stream was interpolated using linear interpolation so that data streams from different sensors used aligned timestamps. Since the

Table 2.1: Merlyn Sensors Used for Modeling

Modality	Sensor
Acceleration (3-axis)	Kionix KX-224-1053 [2]
Ambient Light	ROHM RPR-0521RS [27]
Magnetic Field (3-axis)	ROHM 1422AGMV [24]
Pressure (Barometric)	ROHM BM1383AGLV [28]
Temperature	ROHM BD1020HFV [35]

data were recorded at 16 Hz and reactor operational state changes at a relatively slow rate, linear interpolation is a reasonable approximation for the actual physical value. After this, the temperature and pressure data streams were background corrected using weather data collected from a nearby National Oceanic and Atmospheric Administration (NOAA) [26] facility by subtracting the ambient weather value from the value recorded by the sensor. This mitigated potential confounding trends in the pressure and temperature features related to weather. Then significant outliers were removed from each data stream by eliminating data points larger than four mean average deviations. This was done to eliminate obvious measurement errors (such as an unrealistic measurement of 1000°C by the temperature sensor) which occasionally occur in the data. After this, the x, y, and z components of the magnetometer and accelerometer for each multisensor were combined into a single data stream by taking the L2-norm of the individual coordinate components to get the magnitudes. Then the mean and variance over 10-minute time windows were taken over each data stream. This was a feature engineering step which increased the set of features to include both the average of and variability in the measured data for each sensing modality. Finally, the means and variances were standardized by taking the z-scores of each data point. This transformation was applied so that features with different units of measurement were all put on the same scale. This resulted in 120 features consisting of 60 means and 60 variances from 12 Merlyns over five sensing modalities and 39,745 total samples over the 40-week time-series.

Additionally, the sensors experienced occasional periods of outage due to maintenance, power failures, and faulty components. These outages were treated as data that were missing completely at random [30] for preprocessing purposes. Missing data were filled in with the mean over the entire data stream time-series and an additional feature in the form of a missing flag for each data point set as 1 if the data were missing and 0 otherwise. This resulted in 132 total features consisting of 120 means and variances and 12 missing flags.

Information about nuclear reactor operational state was provided in the form of reactor operator logs. These raw values were used to interpolate reactor power state for each 10-minute window corresponding to the samples on a fill-forward basis. That is, for any given 10-minute window, the interpolated power state is the most recently recorded power state. Since the reactor transitions states rarely over the 40-week time-series as can be seen in Fig. 2.4, fill-forward interpolation best models the true reactor operational state for a given 10-minute window.

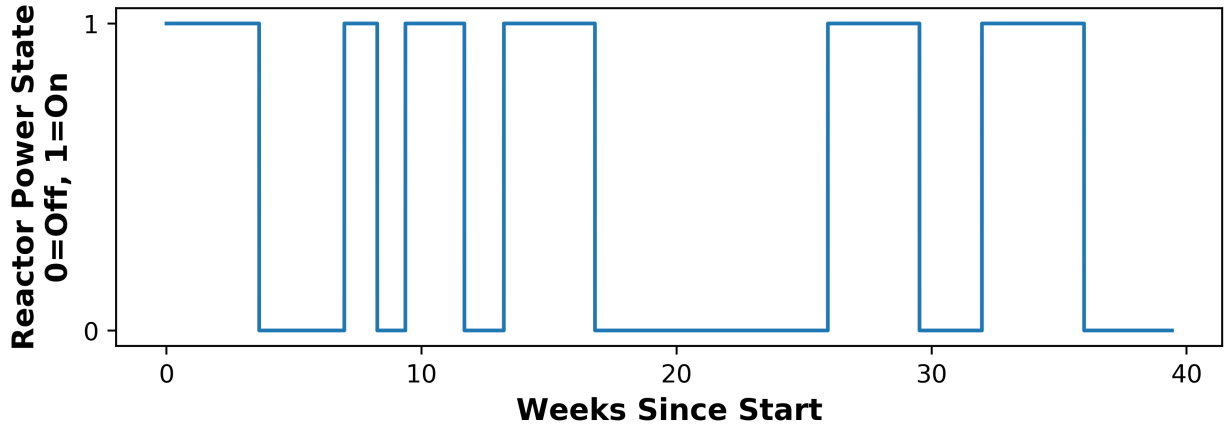

Figure 2.4: Reactor Power State Over the 40-week Time-Series

Given the time-dependent nature of the data, an assumption of independence between samples is incorrect and therefore partitioning the data set into training and testing sets cannot be done with simple random sampling without introducing bias produced by temporal autocorrelation. Instead the method of nested cross-validation [37] was used. In nested cross-validation, the time series is first partitioned into n time segments. These segments are then organized into $n-1$ train/test splits by assigning the ith split to have the $0, ..., i-1$ segments as the train data and the ith segment as the test data set. The training and testing scores for a given model are taken as averages over the $n-1$ splits. A variant of this method used in this work also inserts a buffer between the training and testing partitions to eliminate any bias introduced by temporal autocorrelation across the train and test partitions [29]. While this choice of data partitioning is non-random and can therefore introduce bias in estimating the true training and testing scores, the averages over each split produces estimates which mitigate bias as much as possible [37].

The 40-week time series was split into four approximately equal-sized segments of 10,000 contiguous samples (the 4th segment contains 9,745 samples) which correspond to approximately 10-weeks each. This size was chosen so each train and test partition in each split contained one or more transitions between nuclear reactor power state while keeping the segments similarly sized. Hyperparameter optimization, which would reduce the number of test partitions from three to two for nested cross-validation and therefore increase bias in the estimates of the test scores, was not done. A buffer of one week was put between each train and test partition to mitigate bias introduced by temporal autocorrelation. The approximate partitions for each of the three train/test splits are shown in the illustration in Fig. 2.5. The scores reported in the following sections are averaged over the three splits.

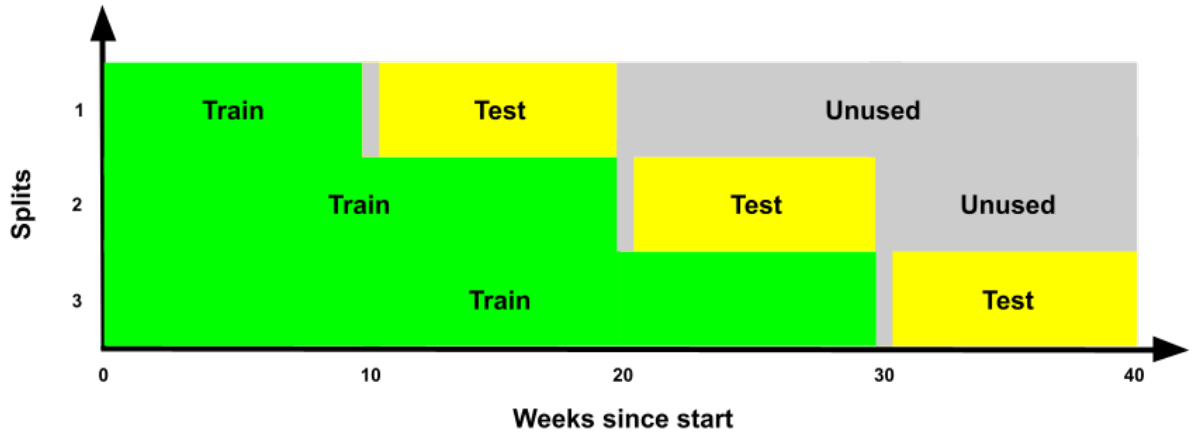

Figure 2.5: Nested Train-Test Split over 40-week Time Series

2.4 Baseline Modeling Efforts

An HMM and a feed-forward neural network were trained and evaluated on the full feature set to get the baseline performances for each model. The HMM used two hidden states to represent the off and on reactor power states and assumes a multivariate Gaussian distribution to model the emissions from each hidden state. The particular implementation of this model comes from the `hmmlearn` package [19]. The feed-forward neural network had an architecture of six hidden layers with 250, 150, 90, 50, 30, 20 units each with ReLU activation, used the Adam optimizer with an initial learning rate of 0.0001, implemented an L1 regularization parameter of 0.001, and ran for 100 epochs. The implementation of this model was done in the `tensorflow` package [3].

The HMM trained on all the features achieved an accuracy of 0.583. The feed-forward neural network trained on all the features achieved an accuracy of 0.811 ± 0.005 ($ci = 0.95$) averaged over 50 runs with randomized initial weights. Plots of the predicted classes versus the actual classes over the three test partitions are shown for the HMM in Fig. 2.6 and for a random chosen run of the feed-forward neural network in Fig. 2.7.

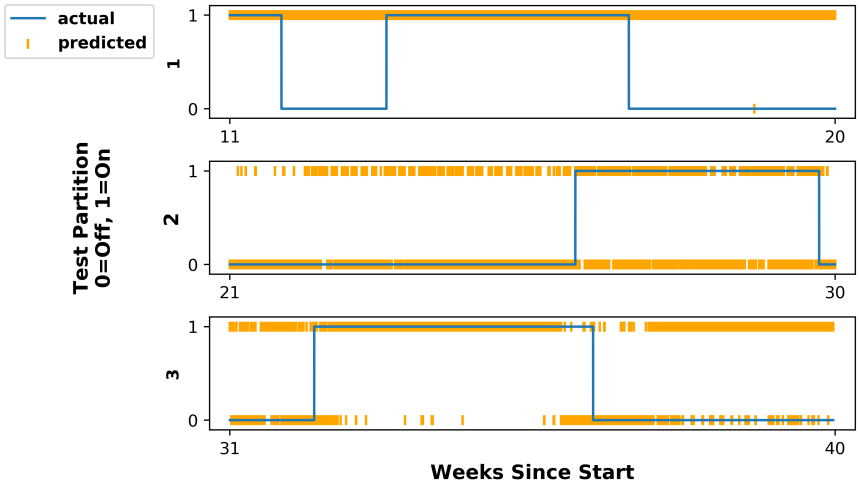

Figure 2.6: Predicted and Actual Reactor Operational State for the baseline HMM

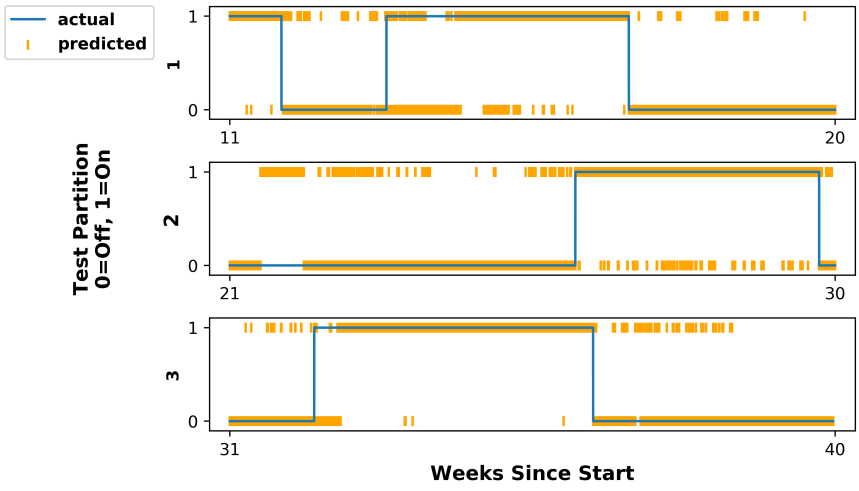

Figure 2.7: Predicted and Actual Reactor Operational State for the baseline Feed-Forward Neural Network Model

Chapter 3

Feature Importance and Wrapper Methods

This work next addresses the question of which nodes and regions at the nuclear facility are the most important for enabling the accurate prediction of nuclear reactor operational state.

3.1 Feature Importance

Node and region importance are closely related to feature importance. Feature importance is the measurement of how much individual features contribute to the overall performance of a model. Feature importance can provide insight into model explainability and help with feature selection by distinguishing between features which do and do not contribute to increased model performance [13]. Feature importance has been used in many studies applying machine learning techniques to data collected by sensors to gain insight into explainability and feature selection. For example, permutation feature importance was used to measure sensor importance in a study classifying sitting posture with sensors to choose a high performing subset of features for a random forest model [39]. Gini importance and backward selection on a k-nearest neighbors model were used to eliminate irrelevant sensors in a study classifying the quality of a laser weld using features derived from sensor data [21]. Permutation feature importance was used to find the optimal placement of sensors on a device [31]. Mutual information-based feature selection and genetic algorithm linear discriminant analysis feature selection were used to determine the most important features derived from sensor data for model-assisted fault detection [17]. From this, it can be seen that measuring the importance of sensors in studies applying machine learning techniques to sensor data is a problem encountered in a wide variety of applications.

3.2 Wrapper Methods

While there are many feature importance methods designed for specific models such as out-of-bag permutation feature importance for random forests [6], SVM-RFE for support vector machines [16], and integrated gradients for neural networks [34], this work focuses on a class of feature importance methods called wrapper methods [22] which 'wrap' around the model to measure the importance of individual features. The key benefit of wrapper methods is that they can be applied to any model in any context to measure feature importance. The specific wrapper methods considered here are Leave One Covariate Out (LOCO) [23] and Forward Feature Selection (FFS) [15]. In LOCO, shown in Algorithm 1, the feature importance of the ith feature is measured as the accuracy difference between a model trained on a full set of features and a model trained on all the features except for the ith feature. In FFS, shown in Algorithm 2, candidate features are iteratively added to a working set of features by greedily adding the candidate feature achieving the highest accuracy to the working set at each iteration. The order in which features are added to the working set provides a ranking of features.

Algorithm 1 Leave One Covariate Out (LOCO)

Ensure: $length(features) > 0$
$importances \leftarrow \{\}$
$base\_score \leftarrow train\_and\_eval(features)$
$i \leftarrow 0$
while $i < length(features)$ **do**
  $ith\_feature \leftarrow features[i]$
  $features\_except\_ith \leftarrow features \setminus \{ith\_feature\}$
  $ith\_score \leftarrow train\_and\_eval(features\_except\_ith)$
  $ith\_importance \leftarrow base\_score - ith\_score$
  $importances[ith\_feature] \leftarrow ith\_importance$
  $i \leftarrow i + 1$
end while
return *importances*

3.3 Node and Region Importance

Node and region importance extend feature importance by directly measuring how much a group of features derived from a node or region, considered together as a group, contribute to the overall performance of a model. While one can measure node and region importance by measuring the importances of their individual constituent features and adding them together, it has been rigorously proven in theory and demonstrated in practice that the importance of a group of features is often different than the sum of the importances of its individual parts [14].

Algorithm 2 Forward Feature Selection (FFS)

Ensure: $length(features) > 0$
$selected \leftarrow []$
$candidates \leftarrow features$
$i \leftarrow 0$
while $i < length(features)$ **do**
  $candidate\_scores \leftarrow \{\}$
  $j \leftarrow 0$
  while $j < length(candidates)$ **do**
    $candidate \leftarrow candidates[j]$
    $feature\_subset \leftarrow selected \cup \{candidate\}$
    $score \leftarrow train\_and\_eval(feature\_subset)$
    $candidate\_scores[candidate] \leftarrow score$
    $j \leftarrow j + 1$
  end while
  $best\_candidate \leftarrow argmax(candidate\_scores)$
  $selected \leftarrow selected + [best\_candidate]$
  $candidates \leftarrow candidates \setminus \{best\_candidate\}$
  $i \leftarrow i + 1$
end while
return *selected*

In fact, the importance of a group of features is the sum of the importances of its individual parts only if the features are uncorrelated which is unrealistic for many practical settings. This problem of correlation is especially relevant for data collected by sensor networks where there are many sources of correlation. For example, there is correlation between different, but related sensing modalities on a single node and between the same sensing modality across two different, nearby nodes. Because of this result, node and region importance provides strong advantages over measuring the importances of individual features and considering the individual importances in a group when evaluating the impact of a given node or region on a classification problem.

While initial efforts in measuring the importance of a group of features have been made for individual models such as random forests [14], multilayer perceptrons [7], support vector machines [40], and least squares regression [38], more generic methods of measuring node and region importance are desirable so they can be applied to any model. This work proposes extending LOCO and FFS, wrapper methods for feature importance, to wrapper methods for node and region importance by grouping features derived from single nodes and spatially collocated sets of nodes. More specifically, LOCO is easily extended to Leave One Node Out (LONO) by measuring the importance of the ith node as the accuracy difference between a model trained on the full set of features derived from the full set of nodes and a model trained on the full set of features except for features derived from the ith node. Leave One

Region Out (LORO), Forward Node Selection (FNS), and Forward Region Selection (FRS) are similarly defined.

Chapter 4

Analysis and Results

Using LONO, FNS, LORO, and FRS analysis, nodes and regions can be categorized into three types: high value, inconsequential, and confounding. High value nodes and regions are found to have an overall positive impact on model accuracy. Inconsequential nodes and regions are found to have no significant positive or negative impact on model accuracy. Confounding nodes and regions are found to have an overall negative impact on model accuracy.

For LONO and LORO analysis the ith node or region is defined as high value if the accuracy of a model trained on all the features minus the accuracy of a model trained on all the features except for those derived from the ith node or region is greater than 0.02. Similarly the ith node or region is defined as confounding if the accuracy difference is less than -0.02. The ith node or region is defined as inconsequential in any other case. For FNS and FRS analysis, the high value nodes and regions are defined as all the nodes or regions selected into the working set before the first decrease in model accuracy. Similarly the confounding nodes and regions are defined as all the nodes selected into the working set after the last increase in model accuracy. All other nodes are regions are defined as inconsequential. While these are arbitrary choices, these definitions of high value, inconsequential, and confounding create a valid mapping between the quantitative values found in each analysis to the qualitative categories. Moreover, the most relevant nodes and regions for further analysis are those which clearly fall into one of the categories rather than those which fall on the edge between high value and inconsequential or inconsequential and confounding. Because of this, the choice of mapping does not matter significantly as long as it is internally consistent and sensible.

Regions were chosen based upon their relationship to particular points of interest at the nuclear facility and their relative distance to other nodes. General descriptions of the defined regions are shown in Table 4.1. The bounding boxes in Fig. 4.1 provide approximate visual cues for each region and should not be interpreted as indicators of the sensing range of the Merlyn multisensors.

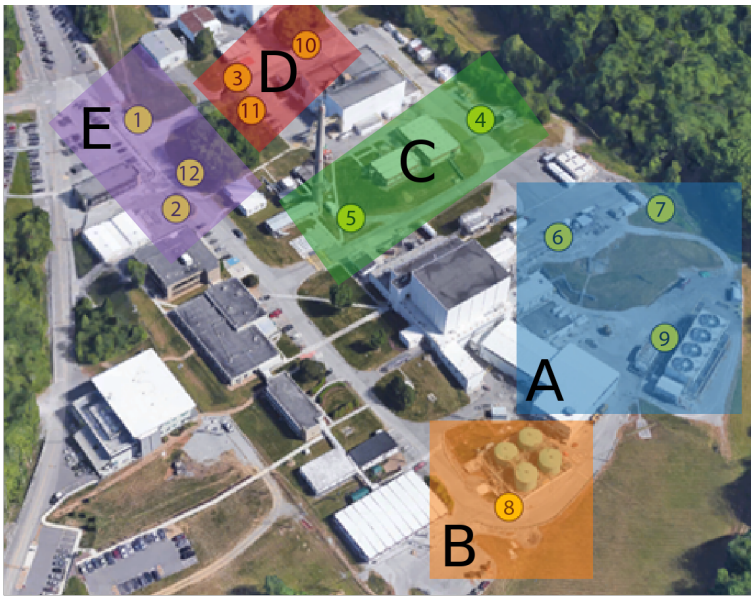

Figure 4.1: Overhead Image of the HFIR Facility With Labeled Nodes And Regions

Table 4.1: Region Descriptions

Region	Nodes	Description
A	6, 7, 9	Reactor Building and Cooling Tower
B	8	Liquid Storage Tanks
C	4, 5	Offices near the REDC Facility
D	3, 10, 11	Target Processing Facility
E	1, 2, 12	Main Entrance to Complex

4.1 Hidden Markov Model

Fig. 4.2 is a plot of the accuracy differences obtained through LONO analysis applied to the HMM. The accuracy differences between the model trained on the full set of features and the model trained on all the features except for those derived from the ith node are ordered from top to bottom in order of highest positive accuracy difference (i.e., most important) to largest negative accuracy difference (i.e., most confounding). From this plot, Node 9 was identified as a high value node due to its strong positive impact on model accuracy. Nodes 2, 5, 7, 4, 1, and 12 were identified as confounding given their strong negative impacts on model accuracy. Node 2 was identified as significantly more confounding than the other confounding nodes.

Fig. 4.3 is a plot of the accuracy differences obtained through LORO analysis applied to the HMM. From this plot, Region A was identified as high value due to its strong positive

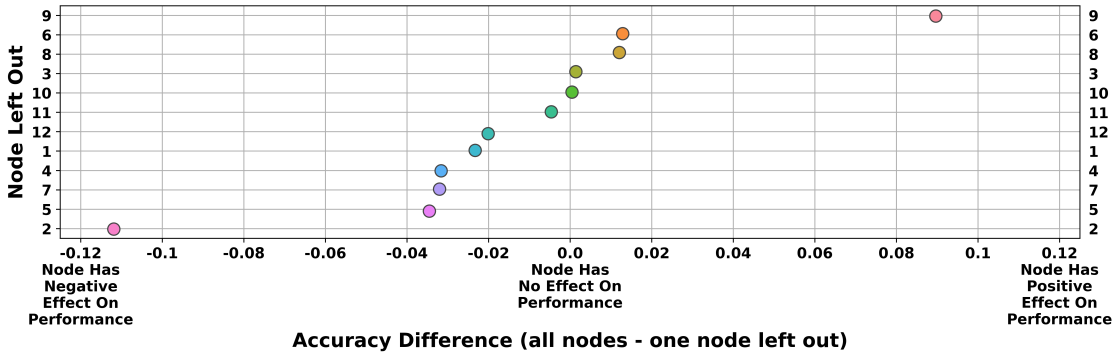

Figure 4.2: LONO Analysis Applied to the HMM

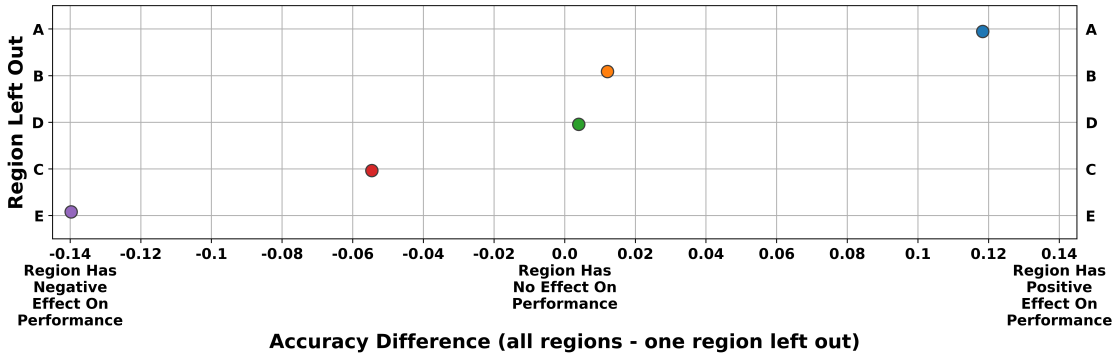

Figure 4.3: LORO Analysis Applied to the HMM

impact on model accuracy. Regions E and C were identified as confounding given their strong negative impacts on model accuracy. Region E was identified as significantly more confounding than Region C.

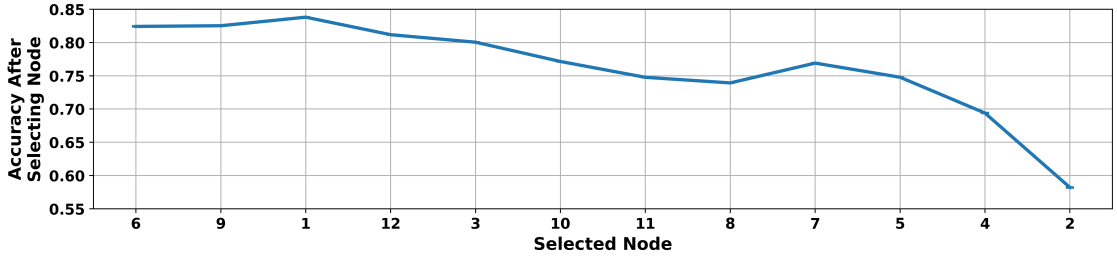

Figure 4.4: FNS Analysis Applied to the HMM

Fig. 4.4 is a plot of the accuracy obtained after adding each node to the working set of nodes during FNS analysis applied to the HMM. The nodes are ordered by importance as determined based on node selection into the working set during the execution of the algorithm (i.e., the node that provides the highest accuracy is selected at each iteration). From this plot, Nodes 6, 9, and 1 were identified as high value nodes due to the accuracy increases after the selection of these nodes into the working set. Nodes 5, 4, and 2 were identified as confounding nodes due to the significant accuracy decreases following the addition of these nodes to the working set.

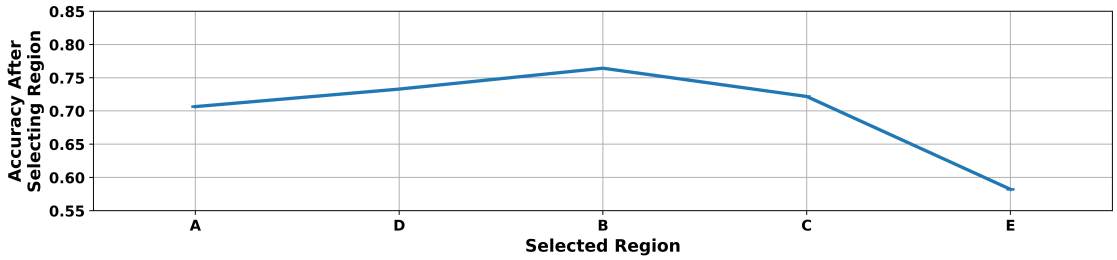

Figure 4.5: FRS Analysis Applied to the HMM

Fig. 4.5 is a plot of the accuracy obtained after adding each region to the working set of regions during FRS analysis applied to the HMM. From this plot, Regions A, D, and B were identified as high value regions due to the significant accuracy increases after the selection of these regions into the working set. Regions C and E were identified as confounding regions due to the significant accuracy decreases following the addition of these nodes to the working set.

These analyses offer valuable insight into inferential hypotheses that could be postulated about HFIR and the HMM produced from the data collected by the sensor network. For example, Nodes 6 and 9 as well as Region A were identified as high value and the corresponding area contains the reactor cooling tower. This points to a potential causal relationship between the cooling tower and reactor operational state. This result is consistent with basic nuclear engineering principles. Nuclear power generation on a MW scale, which is done at HFIR, requires operation of a significant cooling system for the conveyance and removal of heat [18]. The pumps which run this cooling system produce local magnetic fields that may be recorded by the magnetometer and vibrations that may be detected by the accelerometer. The heat rejection into the environment also produces local temperature perturbations that may be sensed by the thermometer. In short, node and region importance analysis combined with knowledge of the HFIR facility and nuclear reactor operations provide justification for a causal relationship between the cooling tower and nuclear reactor operational state.

Additionally, Nodes 2, 12, 4, and 5 and Region C, an area with office buildings, and Region E, the main entrance to the facility, were identified as confounding. These are areas of high foot and vehicle traffic suggesting that foot and vehicle traffic may produce noise

that reduces the accuracy of the HMM. Foot and vehicle traffic produce vibrations that may be recorded by the accelerometer and distortions to the local magnetic field that may be detected by the magnetometer. There is no clear relationship between foot and vehicle traffic and nuclear reactor operational state at HFIR in these areas. This suggests that foot and vehicle traffic may negatively affect the performance of the HMM in its ability to predict nuclear reactor operational state.

These analyses also offer valuable insight relevant to feature selection. The FNS analysis demonstrates that model performance can be increased by only training on features derived from Nodes 6, 9, and 1. An HMM trained only on this subset of nodes resulted in a test accuracy of 0.839 which is a significant improvement over the baseline accuracy (0.583). Plots of the predicted classes versus the actual classes over the three test data partitions for the improved HMM are shown in Fig. 4.6.

An alternative explanation to the foot and vehicle traffic hypothesis comes from the differences in predictions on the test partitions between the baseline and improved HMM. The baseline model transitioned between states (except when it only predicted one class) much more often than the improved model transitioned between states. It is possible that the class conditional probability values $p = P(observation|reactor\ state)$ overwhelmed the values contributed by the transition matrix when calculating the predicted state in the baseline HMM. This result is consistent with results found in naive Bayes classifiers [5] trained on high dimensional data. Because this issue is mitigated as the dimensionality of the data decreases, this is also supported by the result that the removal of some nodes significantly reduced the rate of transitions and increased overall accuracy in the improved HMM. It is also possible that both foot and vehicle traffic and the high dimensionality of the data causing frequent state transitions reduced the performance of the baseline HMM.

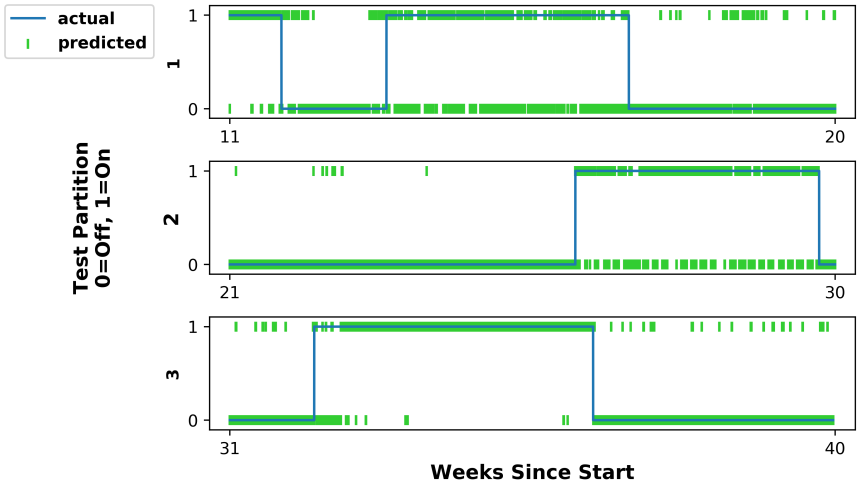

Figure 4.6: Predicted and Actual Reactor Operational State for the Improved HMM

4.2 Feed-Forward Neural Network

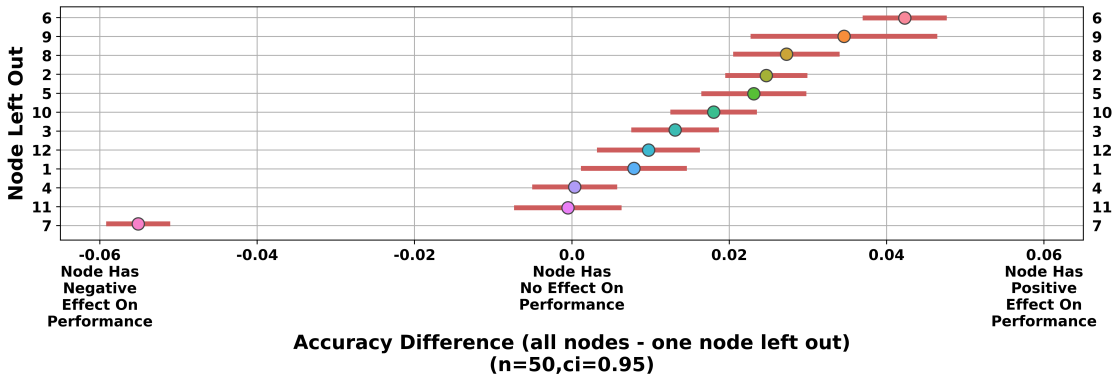

Figure 4.7: LONO Analysis Applied to the Feed-Forward Neural Network

Fig. 4.7 is a plot of the accuracy differences obtained through LONO analysis applied to the feed-forward neural network. For both the baseline full feature set and the feature set with all the features expect for those derived from the ith node, the process of training and evaluating a model was repeated for 50 trials to determine the statistical uncertainty in the assessment. A 95% confidence interval for the accuracy differences for each excluded node was determined. From this plot, Nodes 6, 9, 8, 2, and 5 were identified as high value due to their positive impacts on model accuracy. Nodes 6 and 9 were identified as notably more high value than the other high value nodes. The 95% confidence intervals for nodes 2, 5, and 10 include 0.02 which is the edge between high value and inconsequential. Node 7 was identified as confounding given its strong negative impact on model accuracy.

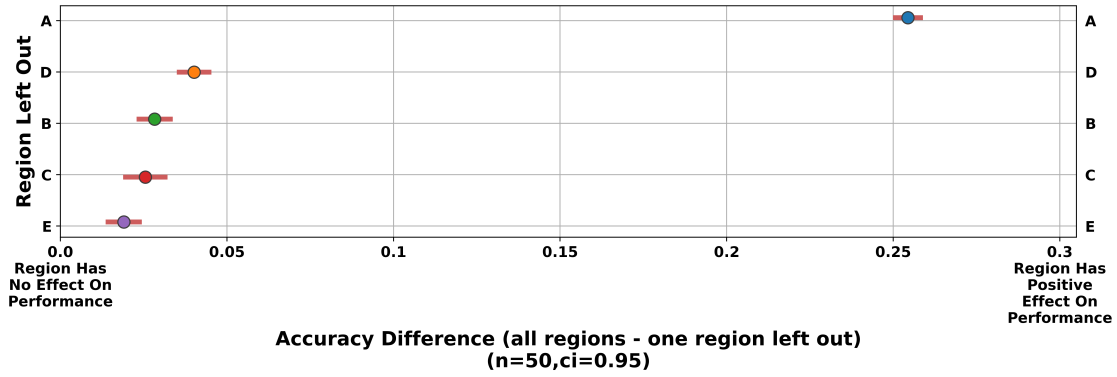

Figure 4.8: LORO Analysis Applied to the Feed-Forward Neural Network

Fig. 4.8 is a plot of the accuracy differences obtained through LORO analysis applied to the feed-forward neural network. From this plot, Regions A, D, B, and C were identified as high value due to their positive impacts on model accuracy. Region A was identified as significantly more high value than the other high value regions. Additionally, the confidence intervals for regions C and E include 0.02 which is the edge between high value and inconsequential. No regions were identified as confounding.

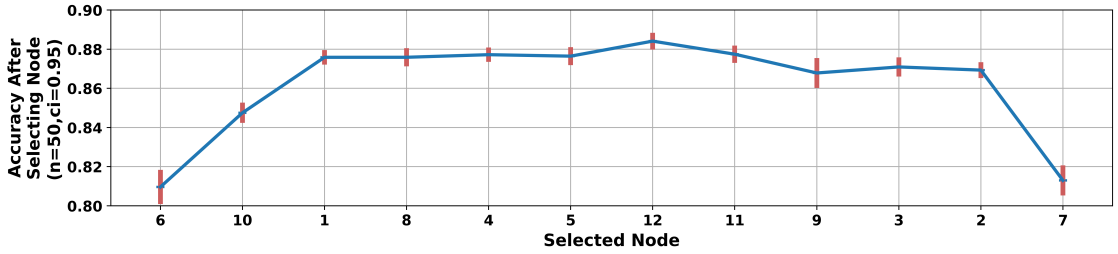

Figure 4.9: FNS Analysis Applied to the Feed-Forward Neural Network

Fig. 4.9 is a plot of the accuracy obtained after adding each node to the working set of nodes during FNS analysis applied to the feed-forward neural network. The performance of a candidate node is taken as the average over 50 trials with randomized initial weights. Additionally, the node selected into the working set is the candidate node with the highest average performance over the 50 trials. The 95% confidence intervals for each selected candidate node are shown in the plot. From this analysis, Nodes 6, 10, 1, 8, 4, 5, and 12 were identified as high value nodes due to the accuracy increases after the selection of these nodes into the working set. Nodes 6, 10, and 1 were identified as significantly more high value than the other high value nodes due to their earlier selection and higher accuracy increases after their addition into the working set. Additionally, the 95% confidence interval for Node 8 includes a decrease in accuracy after adding the node into the working set. Because of this, Nodes 8, 4, 5, and 12 are on the edge between high value and inconsequential. Nodes 2 and 7 were identified as a confounding nodes due to the accuracy decreases following their addition into the working set. Node 7 is significantly more confounding due to the high accuracy decrease after its addition into the working set.

Fig. 4.10 is a plot of the accuracy obtained after adding each region to the working set of regions during FRS analysis applied to the feed-forward neural network. From this analysis, Regions A and D were identified as high value regions due to the accuracy increases after the selection of these regions into the working set. Additionally, an analysis of the first iteration whose scores are shown in Table 4.2 shows that Region A is significantly more high value than Region D. Since the addition of the last region, Region E, resulted in an accuracy increase, Regions C, B, and E were all identified as inconsequential.

Most of these analyses identify Nodes 6 and 9 as well as Region A as high value which is more evidence for a causal relationship between the reactor cooling tower and reactor

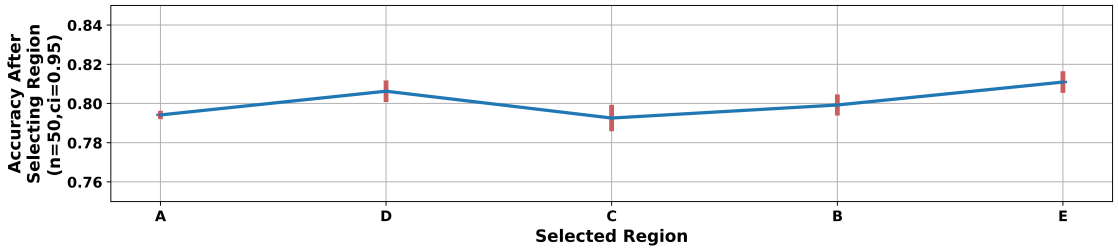

Figure 4.10: FRS Analysis Applied to the Feed-Forward Neural Network

Table 4.2: First Iteration Scores of FRS Analysis

Region	Score
A	0.795 ± 0.002
B	0.556 ± 0.007
C	0.500 ± 0.002
D	0.478 ± 0.002
E	0.586 ± 0.006

operational state. Both LONO and FNS analysis identified Node 7 as the only highly confounding node. This could be due to a component failure on Node 7 which caused a long outage period starting around the 22nd week which extended until the end of the 40-week time-series. In the same way that it has been shown that perturbations in test data sets can dramatically change predictions in adversarial scenarios for neural networks [25], the sensor outage on Node 7 during the evaluation of the 2nd and 3rd test set which affected a significant number of features could have dramatically changed the performance of the model. While this is a limitation of the feed-forward neural network trained on this data, it does nonetheless improve the explainability and interpretability of this model for the purposes of human decision-making.

The FNS analysis demonstrates that model performance can be increased by only training on features derived from Nodes 6, 10, 1, 8, 4, 5, and 12. A feed-forward neural network trained 50 times with different randomized initial weights on this subset of nodes resulted in an average test accuracy of 0.884 ± 0.004 which is an improvement over the baseline average test accuracy (0.811 ± 0.005). Plots of the predicted classes versus the actual classes over the three test data partitions for the improved feed-forward neural network are shown in Fig. 4.11.

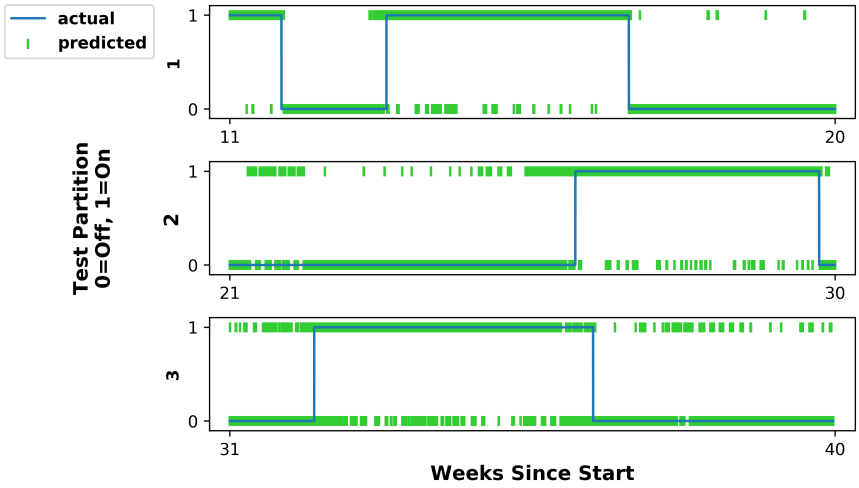

Figure 4.11: Predicted and Actual Reactor Operational State for the Improved Feed-Forward Neural Network

Chapter 5

Conclusion

Node and region importance methods were demonstrated on a problem predicting nuclear reactor operational state using a hidden Markov model and a feed-forward neural network. First, base models were created, then these models were analyzed using node and region importance, and finally the models were improved with feature selection. This allowed for improved understanding of the problem context and predictive models through the postulation of inferential hypotheses about the nuclear facility and resulting models.

The node and region importance methods outlined herein can be applied in any context where sensors are deployed over a spatial area to record data streams used to build predictive machine learning models. Since they are extensions of wrapper methods, they can be applied to any machine learning model whether it be for classification or regression without loss of generality. These methods can be used to identify high value, inconsequential, and confounding nodes and regions which can be used to better understand the problem context, better understand the models generated from the data collected by a sensor network, and improve model performance through feature selection.

Node and region importance helps nontechnical users such as policymakers and analysts better trust the predictions and understand the limitations of an otherwise opaque model applied to sensor network data by providing insight into which nodes and regions drive model performance. It also helps technical practitioners create more accurate models through feature selection. Node and region importance combined with domain knowledge improve the application of machine learning techniques to data collected by sensor networks.

Bibliography

[1] *8-bit AVR Microcontroller with 32K Bytes In-System Programmable Flash*. ATmega328P. Available at https://ww1.microchip.com/downloads/en/DeviceDoc/Atmel-7810-Automotive-Microcontrollers-ATmega328P_Datasheet.pdf, Rev. 7810D-AVR-01/15. Atmel Corporation. Jan. 2015.

[2] *8g / 16g / 32g Tri-axis Digital Accelerometer Specifications*. KX224-1053. Available at https://d10bqar0tuhard.cloudfront.net/en/datasheet/KX224-1053-Specifications-Rev-2.0.pdf, Rev. 2.0. Kionix. Dec. 2017.

[3] M. Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[4] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[5] P. Bennett. "Assessing the Calibration of Naive Bayes' Posterior Estimates". In: (Oct. 2000).

[6] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[7] D. Chakraborty and N. Pal. "Selecting Useful Groups of Features in a Connectionist Framework". In: *IEEE Transactions on Neural Networks* 19.3 (2008), pp. 381–396. DOI: 10.1109/TNN.2007.910730.

[8] G. Coley. *BeagleBone Black System Reference Manual*. BBONE-BLACK-4G. Available at https://cdn.sparkfun.com/datasheets/Dev/Beagle/BBB_SRM_C.pdf, Rev. C.1. Beagleboard. Jan. 2014.

[9] US Department of Energy. *HFIR: Providing unique capabilities for research and isotope production*. Tech. rep. Available at https://www.ornl.gov/file/high-flux-isotope-reactor-fact-sheet/display.

[10] US Department of Energy. *Radiochemical Engineering Development Center*. Tech. rep. Availble at https://www.ornl.gov/file/ised-facility-overview-redc/display.

[11] G. Flynn et al. "Predicting the Power Level of a Nuclear Reactor by Combining Multiple Modalities". In: (May 2019). URL: https://www.osti.gov/biblio/1524353.

[12] Z. Gastelum et al. "Integrating Physical and Informational Sensing to Support Nonproliferation Assessments of Nuclear-Related Facilities". In: (June 2019). URL: https://www.osti.gov/biblio/1641009.

[13] M. Gevrey, I. Dimopoulos, and S. Lek. "Review and comparison of methods to study the contribution of variables in artificial neural network models". In: *Ecological Modelling* 160.3 (2003). Modelling the structure of acquatic communities: concepts, methods and problems., pp. 249–264. ISSN: 0304-3800. DOI: https://doi.org/10.1016/S0304-3800(02)00257-0. URL: https://www.sciencedirect.com/science/article/pii/S0304380002002570.

[14] B. Gregorutti, B. Michel, and P. Saint-Pierre. "Grouped variable importance with random forests and application to multiple functional data analysis". In: *Computational Statistics & Data Analysis* 90 (2015), pp. 15–35. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0167947315000997.

[15] I. Guyon and A. Elisseeff. "An Introduction of Variable and Feature Selection". In: *Journal of Machine Learning Research* 3 (Jan. 2003), pp. 1157–1182. DOI: 10.1162/153244303322753616.

[16] I. Guyon et al. "Gene Selection for Cancer Classification using Support Vector Machines". In: *Machine Learning* 46.1 (Jan. 2002), pp. 389–422. ISSN: 1573-0565. DOI: 10.1023/A:1012487302797. URL: https://doi.org/10.1023/A:1012487302797.

[17] H. Han et al. "Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning". In: *International Journal of Refrigeration* 34.2 (2011), pp. 586–599. ISSN: 0140-7007. DOI: https://doi.org/10.1016/j.ijrefrig.2010.08.011. URL: https://www.sciencedirect.com/science/article/pii/S0140700710001830.

[18] J. Hensley. *Cooling Tower Fundamentals*. 2009.

[19] *HMMLearn*. Available at https://hmmlearn.readthedocs.io/en/latest/. URL: https://hmmlearn.readthedocs.io/en/latest/.

[20] A. Hunter et al. "Artificial Intelligence and National Security". In: (Nov. 2018).

[21] C. Knaak et al. "Machine learning as a comparative tool to determine the relevance of signal features in laser welding". In: *Procedia CIRP* 74 (2018). 10th CIRP Conference on Photonic Technologies [LANE 2018], pp. 623–627. ISSN: 2212-8271. DOI: https://doi.org/10.1016/j.procir.2018.08.073. URL: https://www.sciencedirect.com/science/article/pii/S2212827118308576.

[22] R. Kohavi and G. John. "Wrappers for feature subset selection". In: *Artificial Intelligence* 97.1 (1997). Relevance, pp. 273–324. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(97)00043-X. URL: https://www.sciencedirect.com/science/article/pii/S000437029700043X.

[23] J. Lei et al. "Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111. DOI: 10.1080/01621459.2017.1307116. eprint: https://doi.org/10.1080/01621459.2017.1307116. URL: https://doi.org/10.1080/01621459.2017.1307116.

[24] *Magnetic Sensor Series: 3-Axis Digital Magnetometer IC*. BM1422AGMV. Available at https://fscdn.rohm.com/en/products/databook/datasheet/ic/sensor/geomagnetic/bm1422agmv-e.pdf, Rev.001. ROHM Semiconductor. Oct. 2016.

[25] D. Wagner N. Carlini. "Towards Evaluating the Robustness of Neural Networks". In: (Mar. 2017).

[26] *Climate Data Online*. https://www.ncdc.noaa.gov/cdo-web/. Apr. 2021.

[27] *Optical Proximity Sensor and Ambient Light Sensor with IrLED*. RPR-0521RS. Available at https://fscdn.rohm.com/en/products/databook/datasheet/opto/optical_sensor/opto_module/rpr-0521rs-e.pdf, Rev.001. ROHM Semiconductor. Jan. 2016.

[28] *Pressure Sensor series: Pressure Sensor IC*. BM1383AGLV. Available at https://fscdn.rohm.com/en/products/databook/datasheet/ic/sensor/pressure/bm1383aglv-e.pdf, Rev.003. ROHM Semiconductor. Mar. 2016.

[29] D. Roberts et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In: *Ecography* 40.8 (2017), pp. 913–929. DOI: https://doi.org/10.1111/ecog.02881. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.02881. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881.

[30] D. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (Dec. 1976), pp. 581–592. ISSN: 0006-3444. DOI: 10.1093/biomet/63.3.581. eprint: https://academic.oup.com/biomet/article-pdf/63/3/581/756166/63-3-581.pdf. URL: https://doi.org/10.1093/biomet/63.3.581.

[31] R. Semaan. "Optimal sensor placement using machine learning". In: *Computers & Fluids* 159 (2017), pp. 167–176. ISSN: 0045-7930. DOI: https://doi.org/10.1016/j.compfluid.2017.10.002. URL: https://www.sciencedirect.com/science/article/pii/S0045793017303596.

[32] *SensorShield-EVK-003 Manual*. SensorShield-EVK-003. Available at http://rohmfs.rohm.com/en/products/databook/applinote/ic/sensor/sensorshield-evk-003_ug-e.pdf, Rev.001. ROHM Semiconductor. Apr. 2018.

[33] C. L. Stewart et al. "Multimodal Data Analytics for Nuclear Facility Monitoring". In: *Proceedings of the Institute of Nuclear Materials Management 60th Annual Meeting, Palm Springs*. INMM, 2019, pp. 1024–1034. ISBN: 9781713806110.

[34] M. Sundararajan, A. Taly, and Q. Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

[35] *Temperature Sensor IC*. BD1020HFV. Available at https://fscdn.rohm.com/en/products/databook/datasheet/ic/sensor/temperature/bd1020hfv-e.pdf, Rev.001. ROHM Semiconductor. Nov. 2015.

[36] *Treaty on the Non-Proliferation of Nuclear Weapons*. U.S. Arms Control and Disarmament Agency, Mar. 1970.

[37] S. Varma and R. Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7.1 (Feb. 2006), p. 91. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-91. URL: https://doi.org/10.1186/1471-2105-7-91.

[38] M. Yuan and Y. Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67. DOI: https://doi.org/10.1111/j.1467-9868.2005.00532.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2005.00532.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x.

[39] R. Zemp et al. "Application of Machine Learning Approaches for Classifying Sitting Posture Based on Force and Acceleration Sensors". In: *BioMed Research International* 2016 (Oct. 2016), p. 5978489. ISSN: 2314-6133. DOI: 10.1155/2016/5978489. URL: https://doi.org/10.1155/2016/5978489.

[40] H. Zhang et al. "Variable selection for the multicategory SVM via adaptive sup-norm regularization". In: *Electronic Journal of Statistics* 2.none (2008), pp. 149–167. DOI: 10.1214/08-EJS122. URL: https://doi.org/10.1214/08-EJS122.