Artificial Intelligence based Prediction of Legionella Risk in Drinking Water Systems

Authors: Brian Christopher Dye[1], Vicente Gomez-Alvarez[2], Laura Boczek[2]
[1]US Environmental Protection Agency Region 6 (Dallas)
[2]US Environmental Protection Agency Office of Research and Development (Cincinnati)

ABSTRACT

A neural network model was developed which is capable of yielding an approximate probability of a water sample from a premise plumbing system containing an amount of Legionella pneumophila above the detection limit of 10 MPN per 100 mL. The neural network was trained using the TensorFlow machine learning platform using physico-chemical water quality parameters. Some samples contained additional sample results, such as turbidity and minerals, but these parameters were excluded because the dataset did not contain enough samples with these parameters to obtain a viable result. The parameters were systematically trained using networks of increasing size and complexity before determining the ideal network size for this model.

Initial determination of viability of sample size with various parameters was performed using a neural network composed of the input layer, two hidden layers of 32 nodes with ReLU (Rectified Linear Unit) activation layers, and a final dense layer of one node. Viable parameters were: Total Chlorine, Free Chlorine, Temperature, pH; removed parameters included turbidity and mineral results because the dataset did not contain enough samples with these parameters to obtain a model loss and accuracy outside of the range of error of a positive results correlation.

BACKGROUND

- **State and local government officials issued shelter-in-place recommendations ("social distancing") and recommended the closing or reduced operation of buildings to stop the global pandemic caused by novel coronavirus disease (COVID-19). During this time, unoccupied and low-occupancy buildings might have experienced extended periods of low water demand without proper water management plans (i.e., mitigation). Periods of low or no occupancy can be challenging for building systems and may increase the risk of water system failures and other hazards for occupants.**
- **Reduced consumption of water can cause stagnant water to accumulate in building water systems. Water stagnation can lead to reduced water quality including presence of the bacteria Legionella pneumophila. L. pneumophila is a Gram-negative bacterium and is the major causative agent of Legionnaires' disease.**
- **In the United States, reported cases of Legionnaires' disease have grown by nearly nine times since 2000. Legionella grows best in warm and stagnated water or in building water systems that do not have enough disinfectant to prevent the growth and spread of microbes.**
- **This research investigated the practicality of water quality parameters-based signatures as a screening tool and potential predictor of critical levels of Legionella in a building.**
- **Water quality parameters are relatively inexpensive to measure compared to extensive culture laboratory testing for Legionella. We have developed a convolutional neural network which can be used to predict if a building's plumbing or area of a distribution system is at increased risk of Legionella (i.e., action risk levels).**
- **A machine learning algorithm was developed by designing a neural network utilizing the TensorFlow machine learning platform. The project used data (i.e., physico-chemical water quality parameters) collected by the EPA Office of Research and Development such as temperature, water usage rate, pH, turbidity, disinfectant residual, water source (e.g., taps, pipes, tanks), and detected Legionella threshold cycle levels.**
- **A machine learning algorithm was applied to determine if there is a correlation between these water quality parameters and presence of Legionella.**

REFERENCES, ACKNOWLEDGEMENTS, AND DISCLAIMER

- **The views expressed are his own and do not necessarily represent those of the United States or U.S. EPA.**

- G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.
- Google Brain Team. TensorFlow. Large-Scale Machine Learning on Heterogeneous Distributed Systems.
- J. D. Hunter, 2007, Matplotlib: A 2D Graphics Environment" In Computing in Science & Engineering, Vol. 9, No. 3., pp. 90-95

- My loving wife, Olga Dye, for supporting and encouraging me in my educational pursuits and life

METHODS

- **504 samples were used in the development of the model**
  - **362 for training**
  - **41 used for validation**
  - **101 used for test of post training loss and accuracy**
- **Values were chosen to yield a 90% confidence interval with a 7% margin of error for the test set after removal of the validation set**
- **The optimal network structure for this problem was determined by starting at the smallest network size of no hidden layer, then including one hidden layer of one node then doubling the nodes of the hidden layer with each size increase**
- **Training Neural Network**
  - **Median used 10 training iterations with below parameters but 0.01 training rate and a patience of 100 epochs**
  - **Binary cross-entropy**
  - **Adam optimizer with a training rate of 0.001**
  - **Patience of 10 epochs**
  - **Batch size of a full epoch**
  - **3205 epochs and yielded**
  - **Training accuracy 86.19%**
  - **Training loss 0.2635**
  - **Validation accuracy 87.80%**
  - **Validation loss 0.2726**
  - **Test accuracy 86.14%**
  - **Test loss 0.2832**

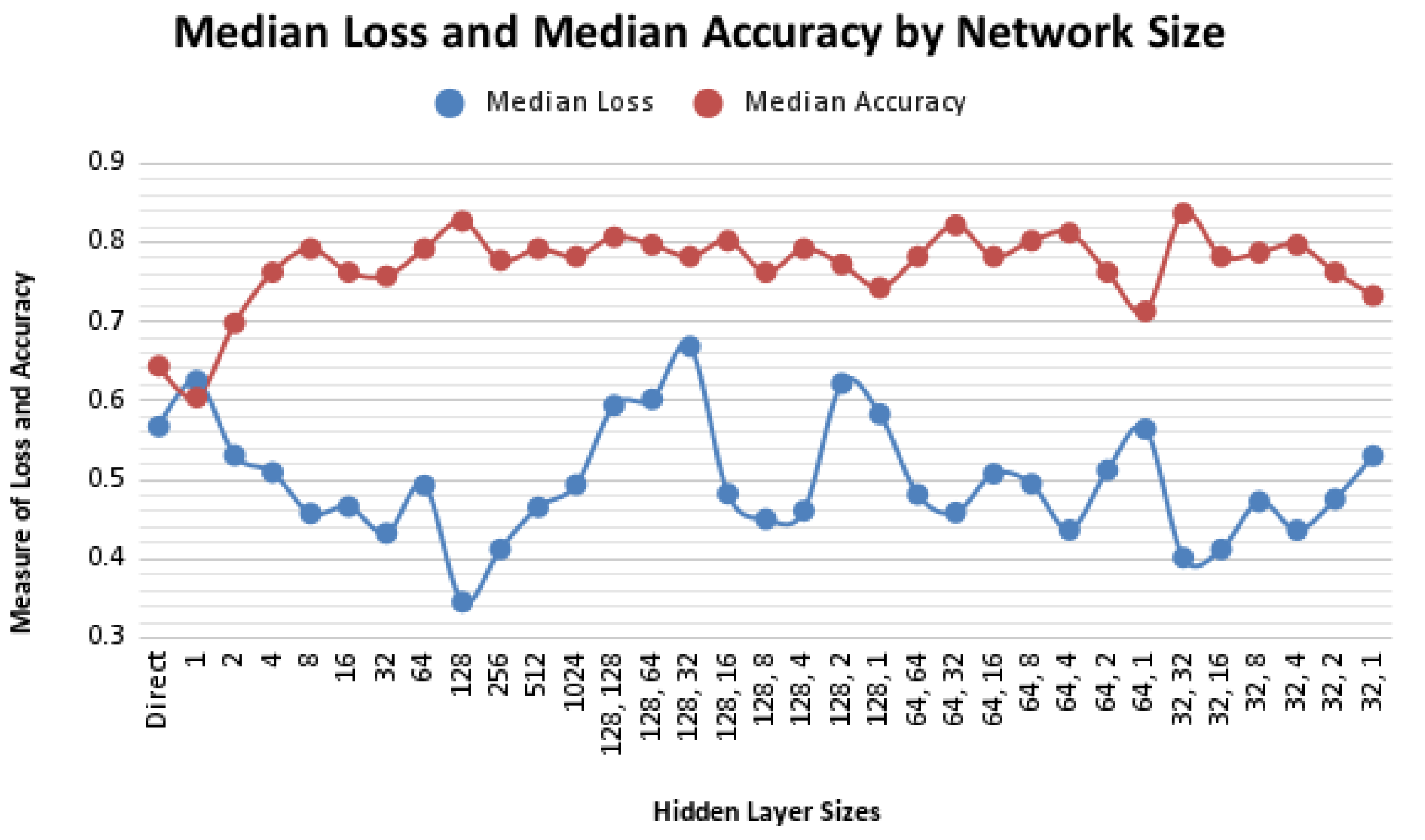
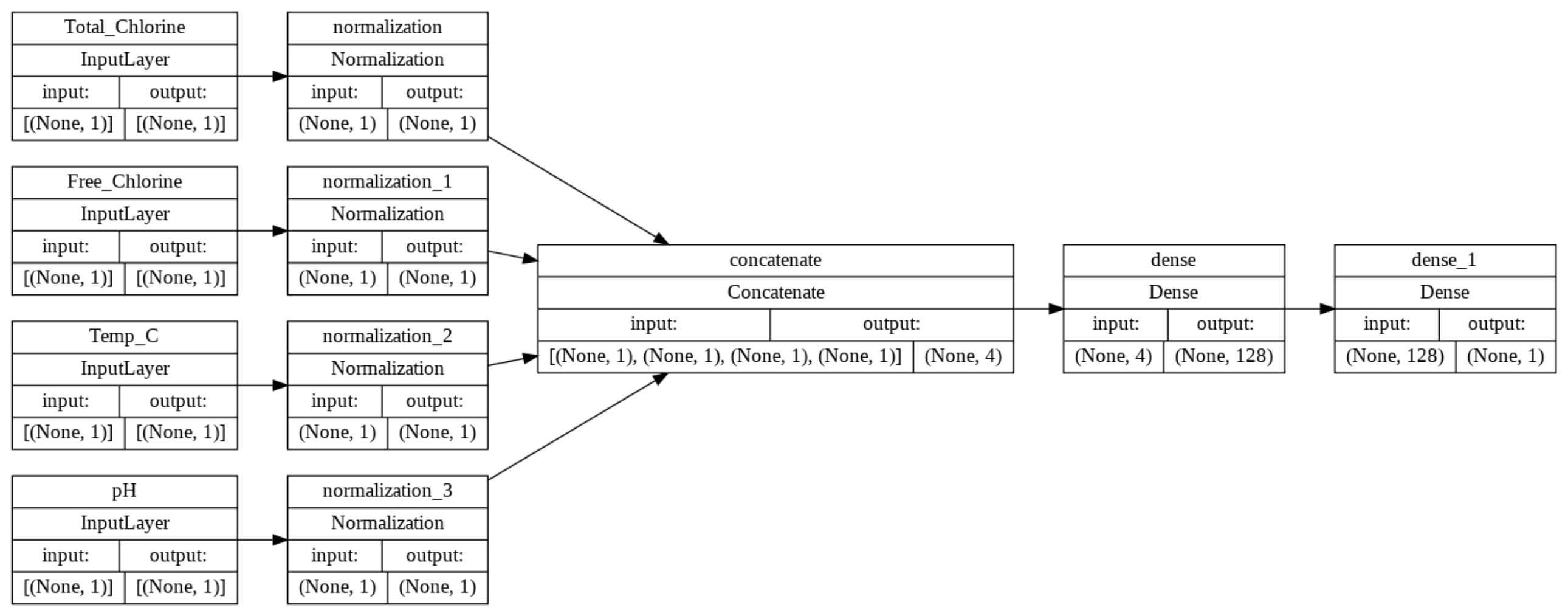


Fig. 1 - Measure of loss and accuracy of various neural network structures. Direct is a network composed of only the input and output layer. The values following along the x-axis are the number of nodes in each hidden layer. The layers with two numbers are represented by the first number being the first hidden layer following input, followed by the number of nodes in the second hidden layer following input.



Total_Chlorine
InputLayer
input: [(None, 1)] output: [(None, 1)]

Free_Chlorine
InputLayer
input: [(None, 1)] output: [(None, 1)]

Temp_C
InputLayer
input: [(None, 1)] output: [(None, 1)]

pH
InputLayer
input: [(None, 1)] output: [(None, 1)]

normalization
Normalization
input: (None, 1) output: (None, 1)

normalization_1
Normalization
input: (None, 1) output: (None, 1)

normalization_2
Normalization
input: (None, 1) output: (None, 1)

normalization_3
Normalization
input: (None, 1) output: (None, 1)

concatenate
Concatenate
input: [(None, 1), (None, 1), (None, 1), (None, 1)] output: (None, 4)

dense
Dense
input: (None, 4) output: (None, 128)

dense_1
Dense
input: (None, 128) output: (None, 1)

Fig. 2 - Graphical chart of the four input parameters through normalization, concatenation, a densely connect 128 node ReLU activated hidden layer, then to a densely connected 1 node sigmoid activated output layer.

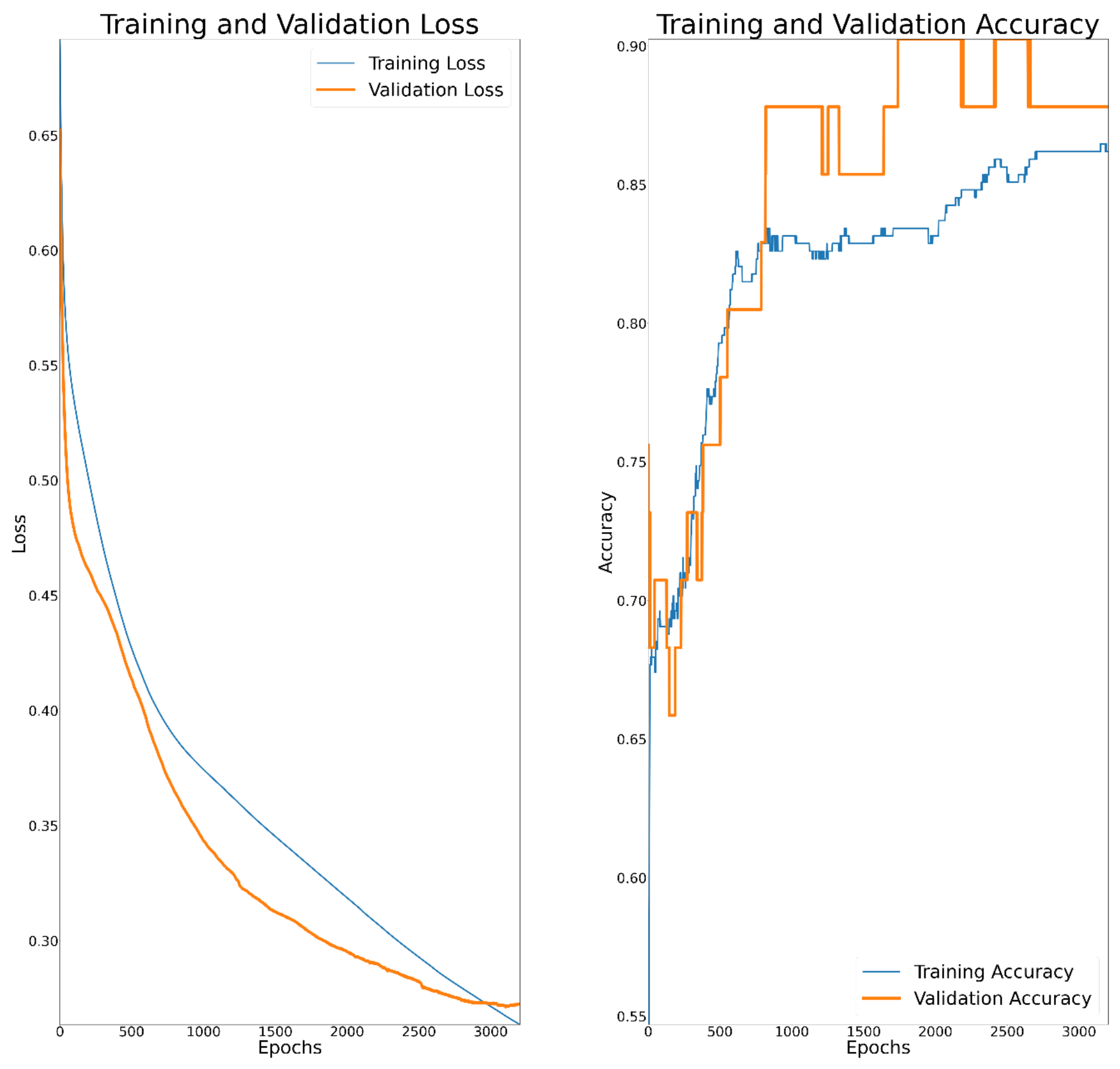


Fig 3 – Above Left: loss of training and validation datasets; validation flattens near end of training epochs as determined by a patience of 100. Above Right: accuracy of training and validation datasets; validation has larger intervals along the y-axis because of the smaller dataset used with validation (41 samples) than training (101 samples).

RESULTS

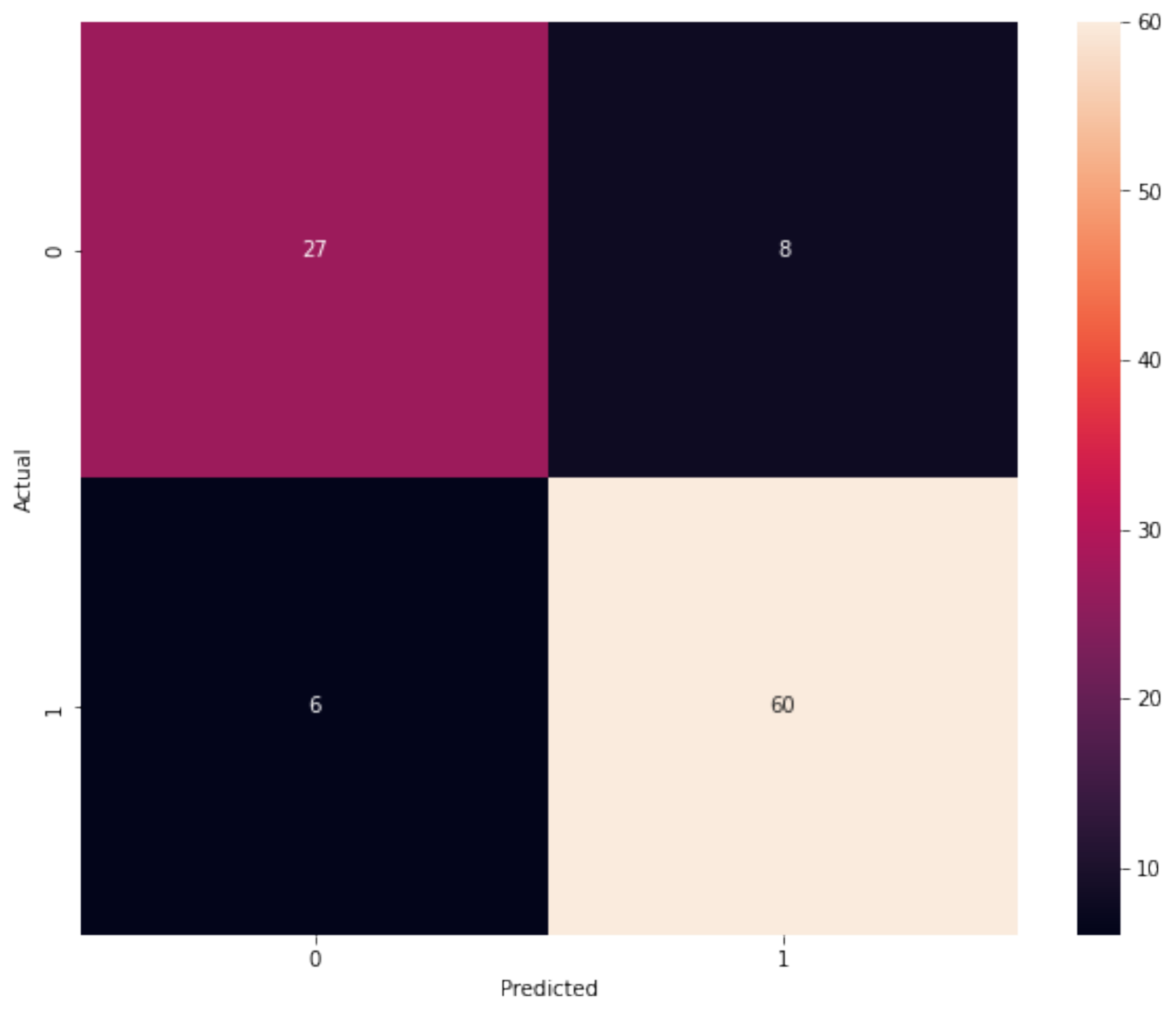


Fig. 4 - Confusion Matrix displaying results from Test dataset of samples. 1 represents a positive detect as a probability of 50% or more. 0 represents a negative detect.

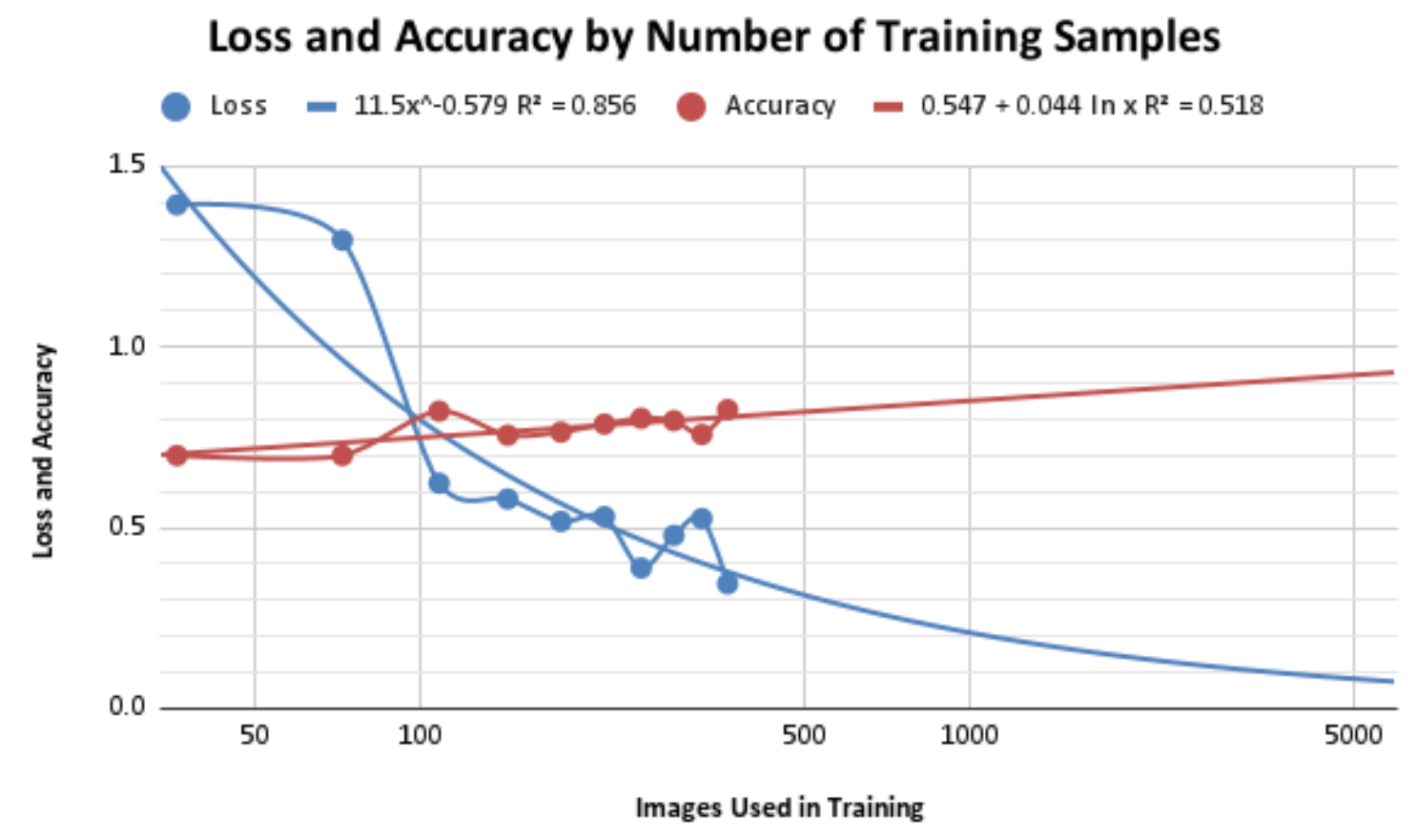


Fig. 5 - Logarithmic graph of model performance after training neural network with ten training sets start at 10% the size of the original training dataset and continuing to increase in intervals of 10% until reaching the size of the full training dataset. Trend lines have been graphed to project performance of the neural network when trained with increasing amounts of samples.

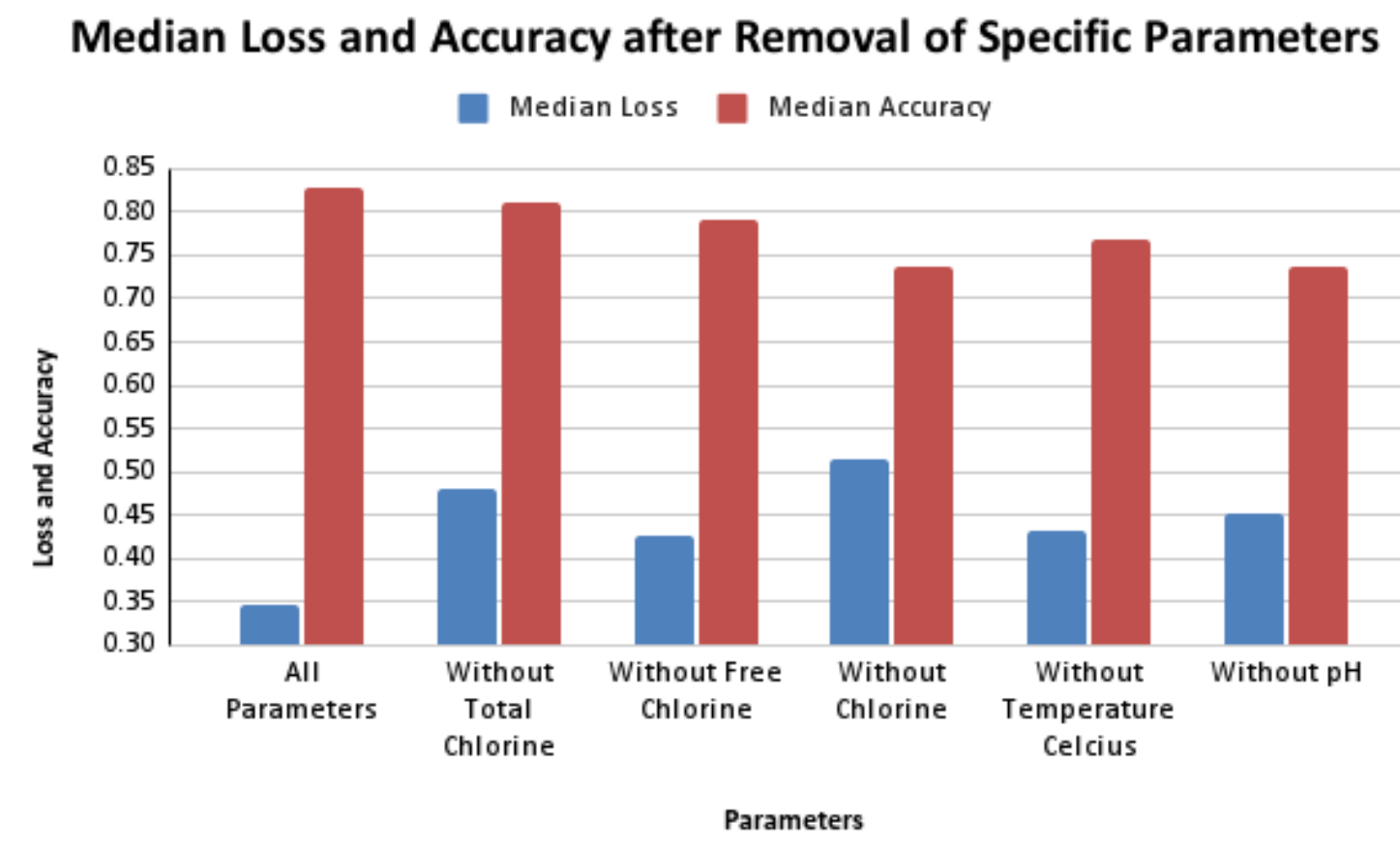


Fig. – 6 Comparison of model performance after removal of individual parameters and combined chlorine parameters. "All Parameters" is the performance of the model without the removal of any parameters.

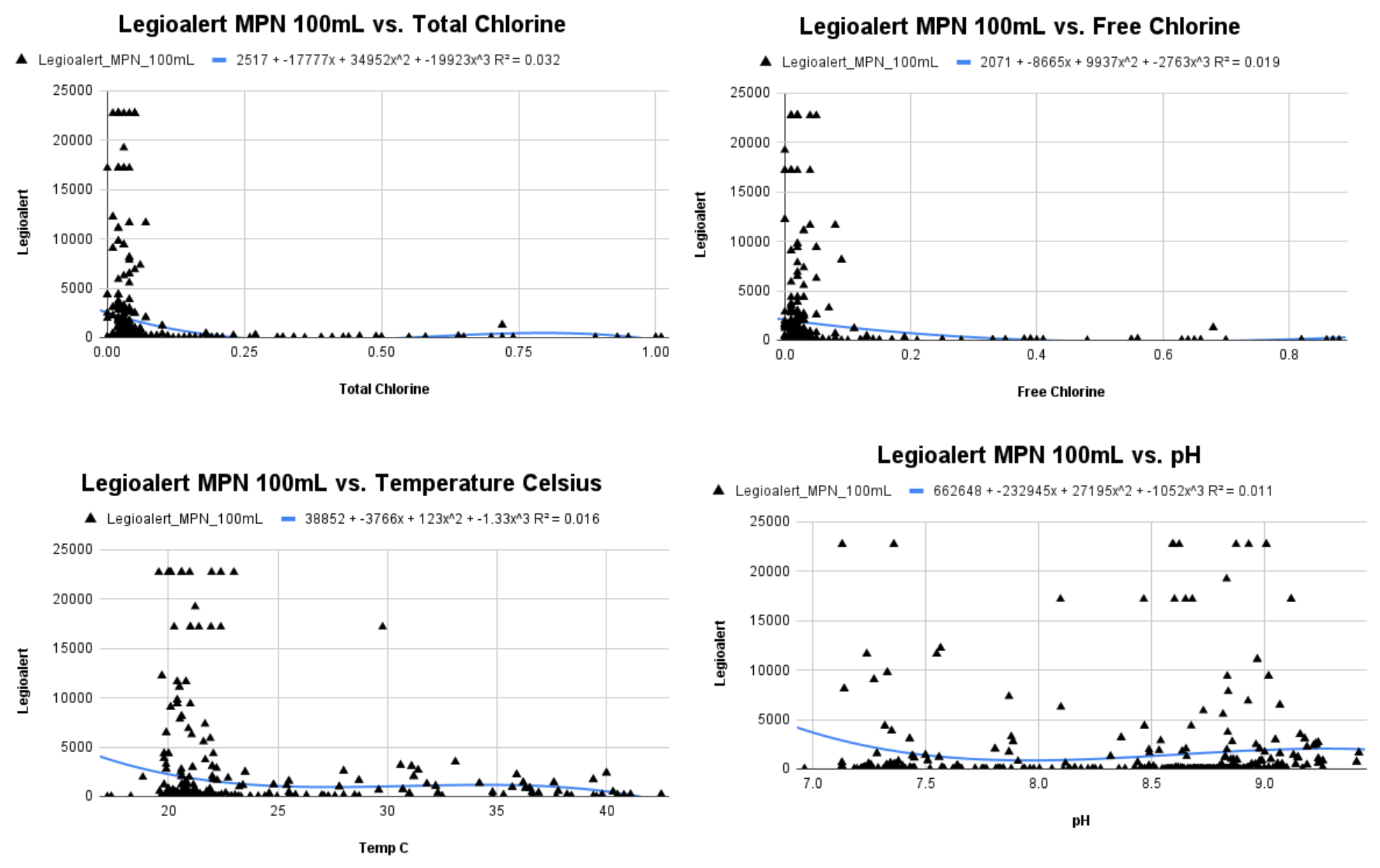


Fig. 7 - Levels of Legionella detected vs: Total Chlorine, Free Chlorine, Temperature Celsius, and pH.

CONCLUSIONS

- **Neural Network was able to identify parameters with the highest correlation for detection of *Legionella***
- **Used parameters: Total Chlorine, Free Chlorine, Temperature, pH**
- **Combined parameters were able to predict whether a sample of water would have detectable levels of *Legionella***
- **Recall = 0.91**
- **Precision = 0.88**
- **Achieved an F1 score of 0.90**

Further research will focus on the development of a monitoring system that may be used to alert building managers and water system operators of an increased risk of *Legionella*, leading to either testing or mitigation efforts.

Knowing the Legionella risk in a premise water system is useful when determining if building flushing is necessary, such as after periods of stagnation or where water may be aerosolized. Specific levels of Legionella were not able to be determined with the number of samples used in this research but may be possible with more sample results. It may be possible to use neural network-based machine learning to determine other water quality parameters quickly and inexpensively. Timely determination of water quality risks is necessary to immediately realize or mitigate potential risks while waiting for laboratory results.