# Iowa Initiative for Artificial Intelligence

# Final Report

| | |
|---|---|
| Project title: | Using Machine Learning to Predict Normocalcemic Primary Hyperparathyroidism |
| Principal Investigator: | Elizabeth Bowen; Nitin Pagedar; Zachary Urdang |
| Prepared by (IIAI): | Yanan Liu |
| Other investigators: | |
| Date: | |

| | |
|---|---|
| Were specific aims fulfilled: | Y |
| Readiness for extramural proposal? | N |
| If yes … Planned submission date | |
| Funding agency | |
| Grant mechanism | |
| If no … Why not? What went wrong? | Dataset is too small, PI will get more data in the future and test this pipeline. |

**Brief summary of accomplished results:**

We have developed and validated a Random Forest model to accurately predict positive/negative control patients with normocalcemic primary hyperparathyroidism using identified variables that will be commonly available in patient charts. These variables include: serum calcium and phosphorus, vitamin D, creatinine, systolic and diastolic blood pressure, age, BMI, weight, race, history of osteoporosis medications, and ICD-10 codes indicating a history of kidney stones, osteoporosis, osteopenia and/or fractures. The best prediction accuracy is 79% and F1 is 0.81.

**Research report:**

**Aims (provided by PI):**

We propose to address uncertainties in diagnosing and managing this condition by developing a machine learning algorithm to assist in recognizing patients with normocalcemic primary hyperparathyroidism.

**Data:**

Given the low prevalence of this condition in the population, we started with general population 100k patient data from the TriNetX database.

Inclusion criteria for positive controls: age 18 or older, primary hyperparathyroidism as an ICD-10 diagnosis code in the chart, history of parathyroidectomy.

Exclusion criteria for positive controls: age <18, history of hypercalcemia defined as calcium level of >=10.5 in the chart, or documentation suggesting history of hypercalcemia.

Inclusion criteria for negative controls: age 18 or older, has had parathyroid hormone and calcium checked in the past.

Exclusion criteria for negative controls: any history of parathyroid surgery, any history of a diagnosis containing the word “parathyroid”, any history of hypercalcemia or hyperparathyroidism of any type, any history of hypocalcemia, any history of hypoparathyroidism, any history of calcium >=10.5 mg/dL or parathyroid hormone >=65 pg/mL.

We have run the relevant inclusion/exclusion criteria through 100k data: 23 positive controls meet the criteria, and 559 negative controls meet the criteria.

**AI/ML Approach:**

In this study, supervised machine learning algorithm was implemented for prediction using Python. As many extracted features may be noisy, or highly correlated with each other, Random Forest (RF) algorithm was selected to predict responder/non-responder group. A random forest classifier constructs multiple decision trees using a random subset of data. Each tree makes a prediction, and the final output is determined by integrating the individual predictions. Random forest also helps reduce overfitting and can provide feature importance rankings.

To help boost the model’s performance on a small dataset, K-Fold Cross Validation was used. This technique divides the dataset into k subsets or folds. The goal is to make sure every data point gets a chance to be in both the training and testing sets. On the first iteration the model is trained on k-1 folds and then tested on the remaining fold. On the second iteration a different fold is used for testing and the process is repeated for the k number of iterations. This method helps estimate the model’s performance on a limited data set.

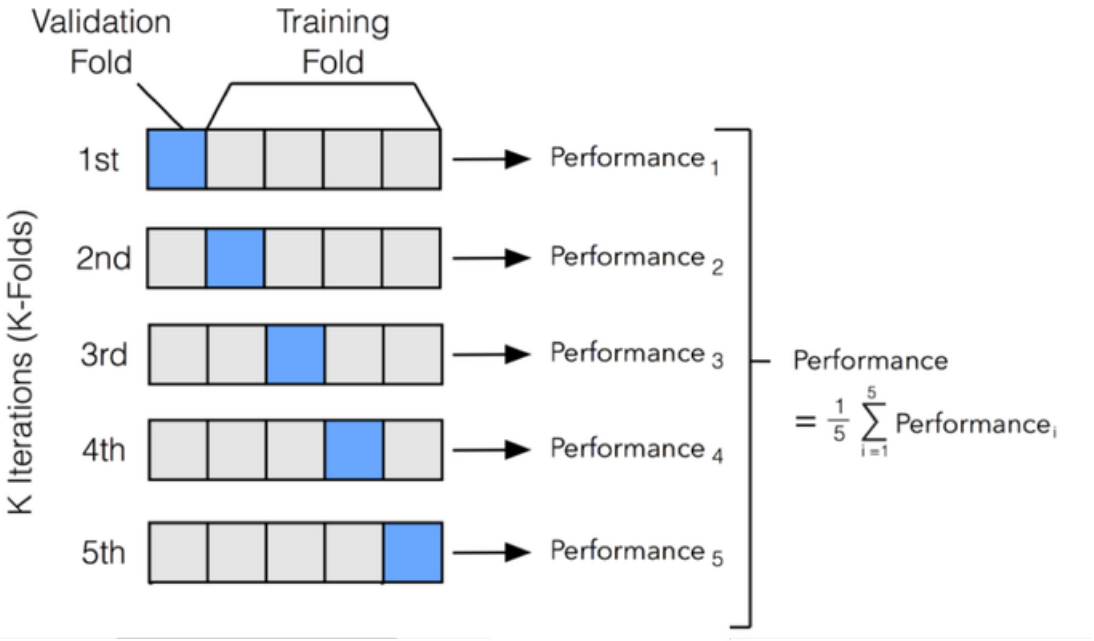


Figure 1. Representation of 5-fold cross-validation technique

**Results:**

The original dataset is unbalanced dataset. The accuracy is 96% and F1 is 0. Confusion matrix is [559,0, 23,0]. By creating a balanced dataset (23 positives and 23 negatives), the prediction accuracy is 79% and F1 is 0.81. Confusion matrix is [15,8, 2,21].