Failure Sources in Machine Learning for Medicine—A Study

Hana Ahmed, Roselyne Tchoua, Jay Lofstead

2nd Workshop on E-science ReseaRch leading tO negative Results (ERROR '22)

October 10, 2022

Salt Lake City, Utah, USA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.

SAND: 1630266

Reproducibility in medical and biomedical research

Reproducibility: "obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis" (NASEM)

- "Reproducibility crisis" in scientific fields (Baker, 2016).
- Lack of transparency regarding:
  - Protocols and raw data
  - Funding sources, potential conflicts of interest
- Open science

AI/ML applications to medicine should meet the same standards expected of medical research.

Trustworthy ML

ML models should be reproducible to be considered trustworthy.

For ML, reproducibility means:

- Ensuring that ML models can be regenerated with identical accuracy and transparency.
- Managing the factors that cause variance in model performance and quality (e.g., pseudo-random numbers, training and testing data, etc.)

This paper identifies challenges to reproducibility in the model design, testing, and publication stages of ML methods for medical data sets.

Source Publications

1. “A comparison of machine learning algorithms for diabetes prediction” (Khanam et al., 2021)
2. “Machine Learning-Based Prediction Models of Coronary Heart Disease Using Gaussian Naïve Bayes and Random Forest Algorithms” (Bernando et al., 2021)

Strengths:

- Robust comparison of results using different ML types and architectures
- Detailed structural and training information
- Detailed data preprocessing instructions

Failure sources:

- Evaluating models on selected subsets of data
- Data preprocessing methods and tools
- Missing information in source publication

Experiments

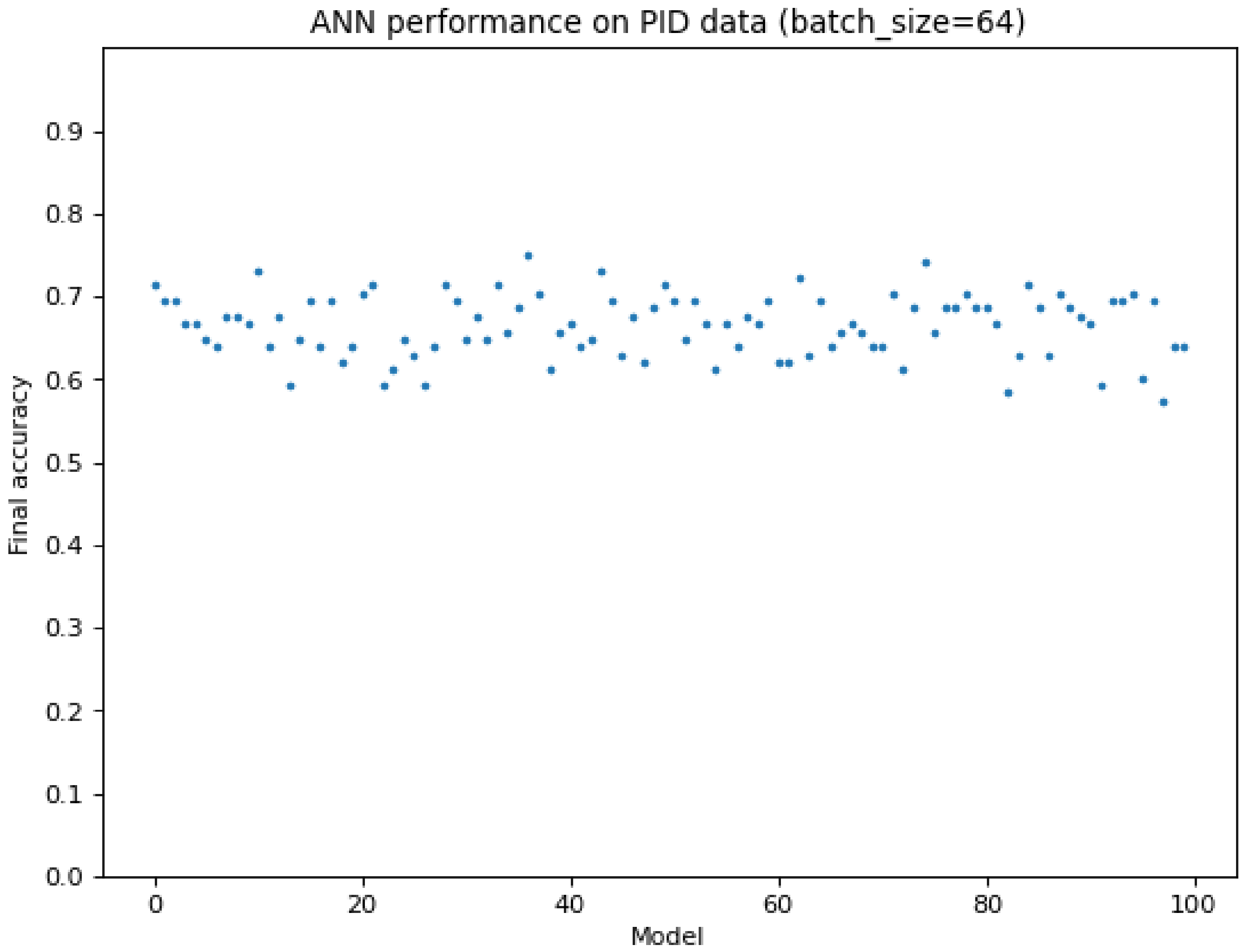
Experiment 1: Predicting diabetes with ANNs

- Replicating an **Artificial Neural Network** evaluated on the **Pima Indian Diabetes (PID)** data set
- 88.6% accuracy reached in published study (Khanam et al., 2021)
- Model specifications
  - **Structure**: 4 dense layers, binary output (to indicate diagnosis)
  - **Learning rate:** .01
  - **Epochs:** 400
  - **Train/test split:** 85/15
  - **Batch size: unspecified***
  - **PRNG seed: unspecified**

Data preprocessing: PID

- WEKA data mining
- Discrepancies between our results and source paper
  - 0 missing values, but 652 reported
  - 49 outliers, but 45 reported
  - 0 extreme values, 26 reported
  - 719 remaining instances, 699 reported
- Feature selection using Pearson correlation coefficient
- Normalization*

Results 1/3

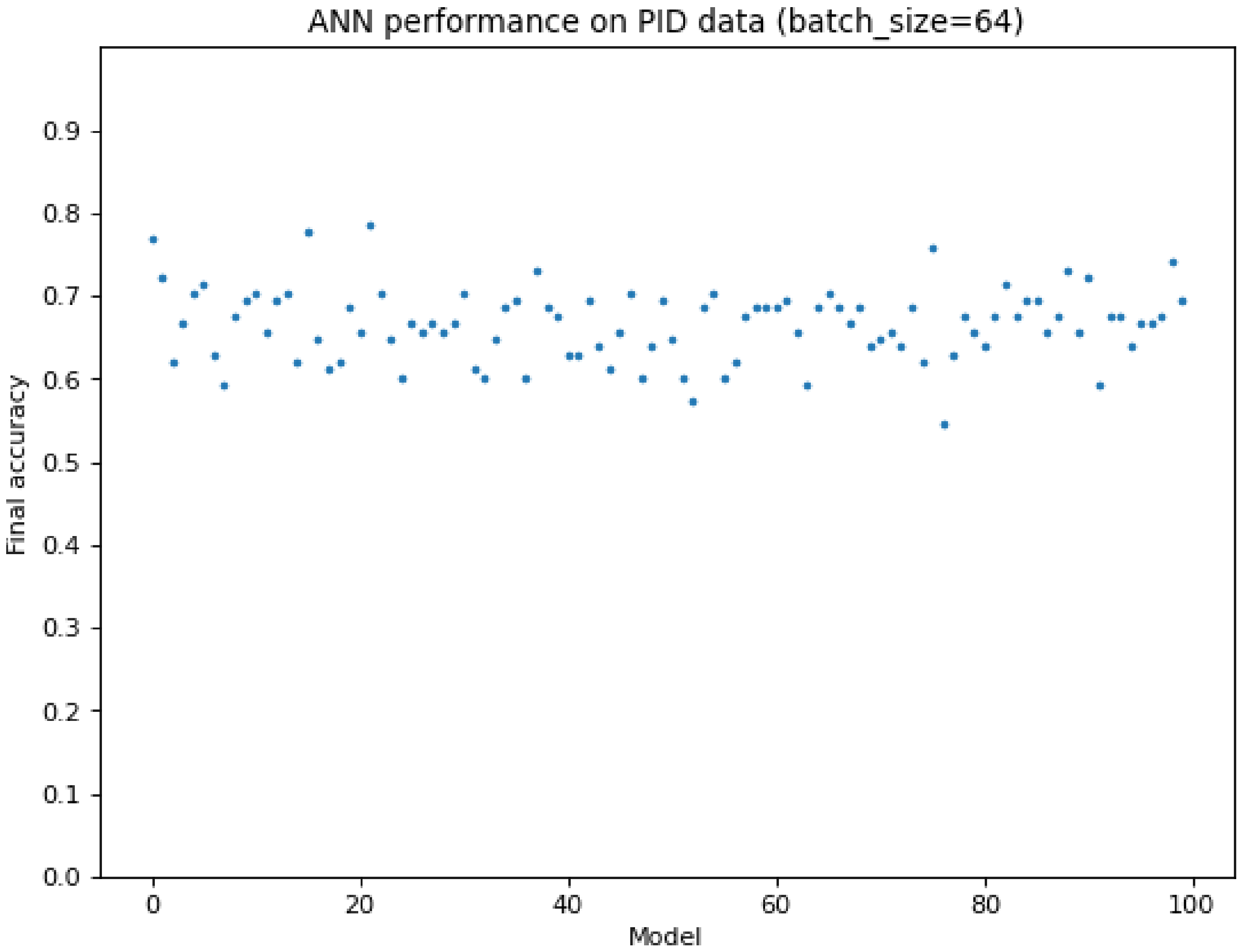


- Batch size set to 64
- PRNG seed varied using datetime function
- Un-normalized data

Best model: 75%
Worst model: 57.41%

Did not reach target of 88.6%

Results 2/3

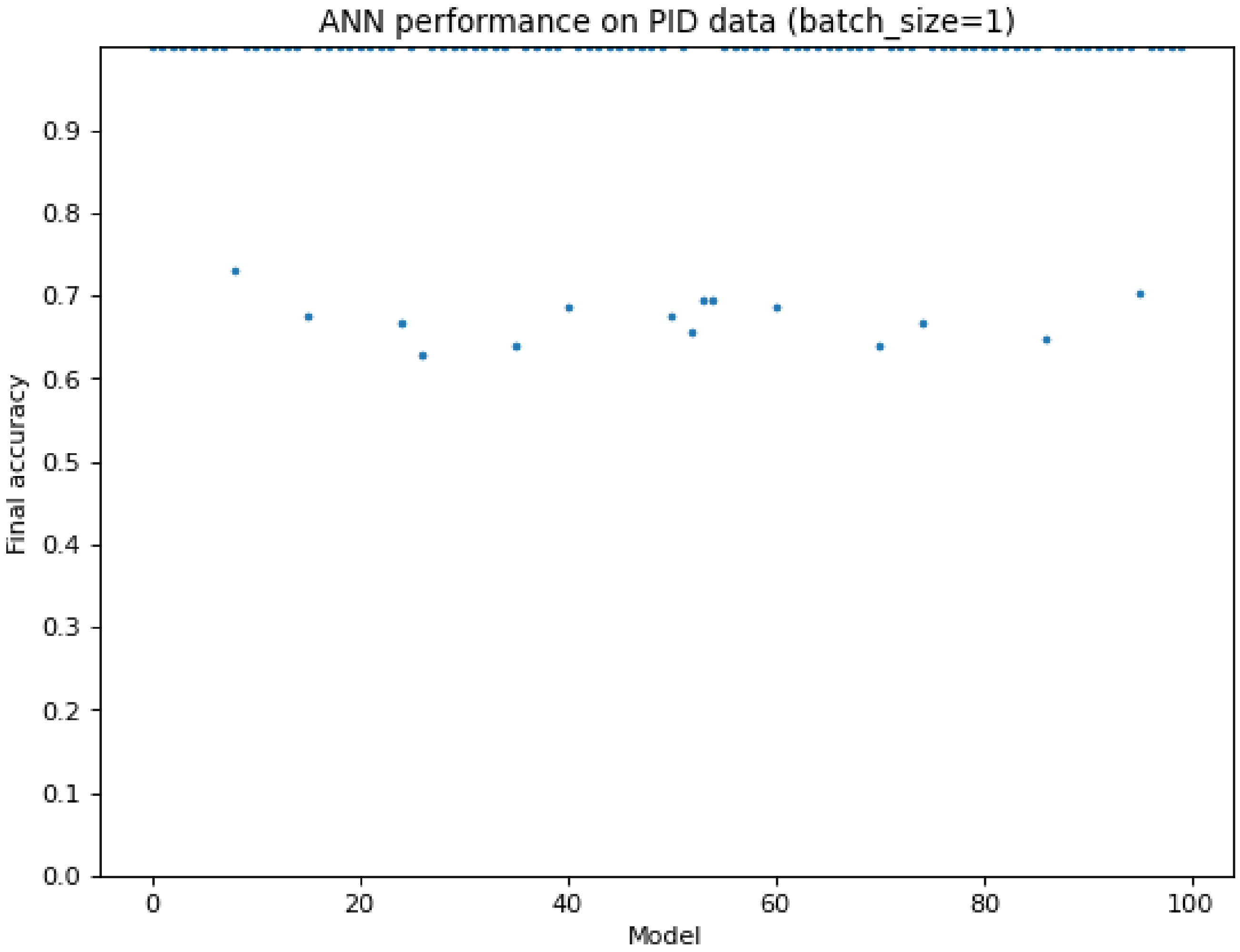


- Batch size set to 64
- PRNG seed varied using datetime function
- Normalized data

Best model: 78.7% accuracy
Worst model: 54.63% accuracy

Did not reach target of 88.6%

Results 3/3



- Batch size set to 1 **after correspondence with authors**
- PRNG seed varied using datetime function
- Normalized data

Best model: 100% accuracy
Worst model: 62.96%

Surpassed 88.6%, but never replicated the target accuracy

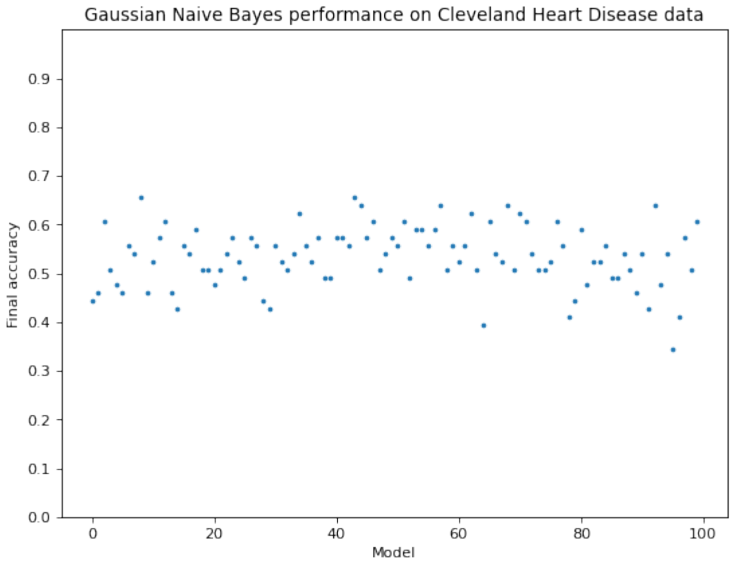
Experiment 2: Predicting heart disease using Naïve Bayes & Random Forest classifiers

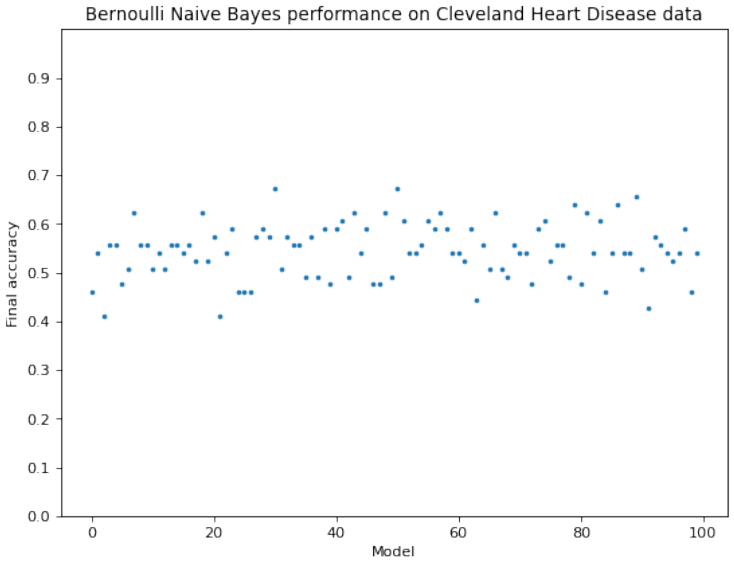
- Replicating **Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and Random Forest classifiers** evaluated on the **Cleveland Heart Disease** data set
- 85%, 85%, and 75% accuracy reached, respectively, in published study (Bernando et al., 2021)
- Model specifications:
  - Python Scikit-learn default parameters
  - Train/test split: 80/20

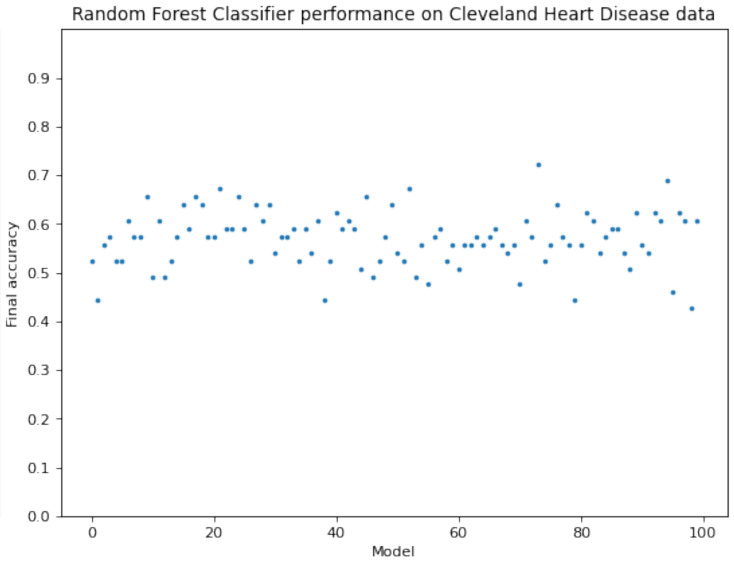
Data preprocessing: Heart Disease

- 4 Heart Disease data sets: Cleveland, Switzerland, Hungary, and Long Beach
  - Source publication only uses Cleveland to train/test models
- No instructions for missing values, outliers/extreme values, etc.
  - Replaced missing values with the mean of the column values
  - Un-normalized data

Cleveland Heart Disease

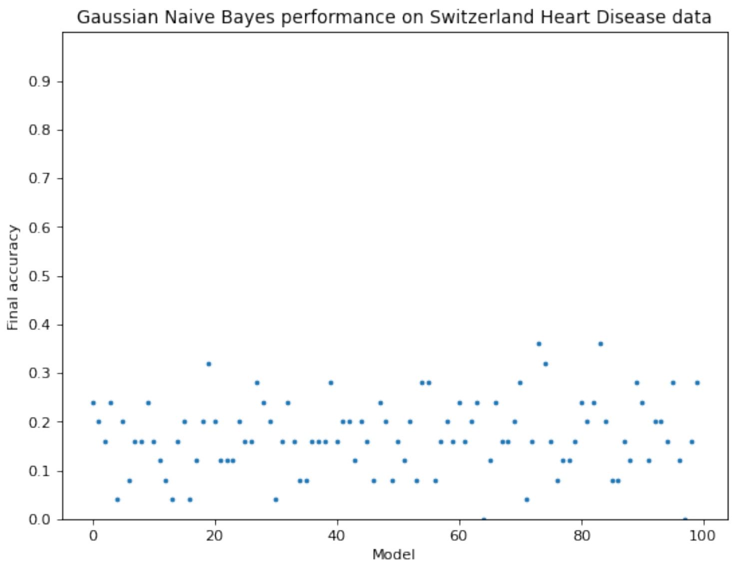


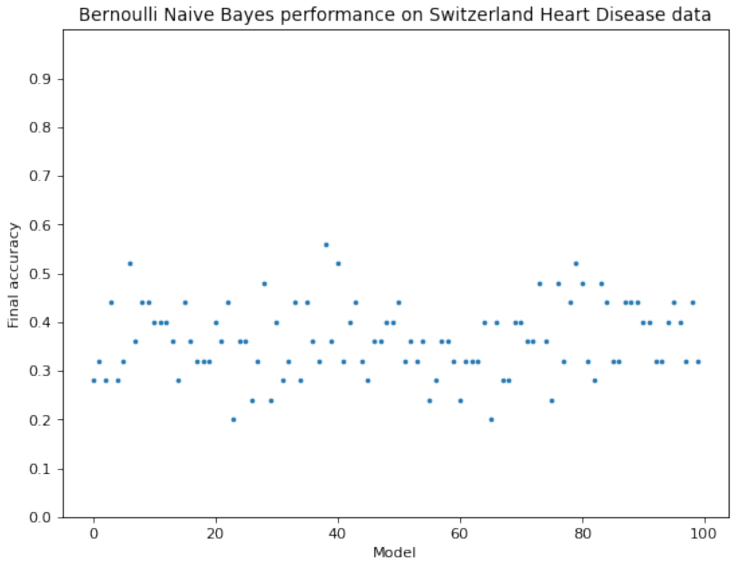


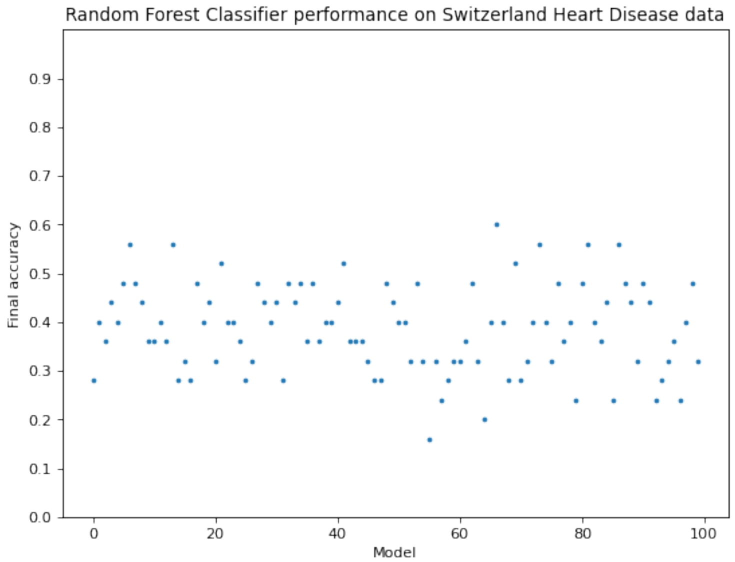


- 100 models each
- Best model accuracies
    - GNB: 65.57%
    - BNB: 67.21%
    - RF: 72.13 %
- Original results: 85%, 85%, 75% (respectively)
    - **Model accuracies were not reproduced**
    - **Best model from our samples was a Random Forest classifier**

Switzerland Heart Disease

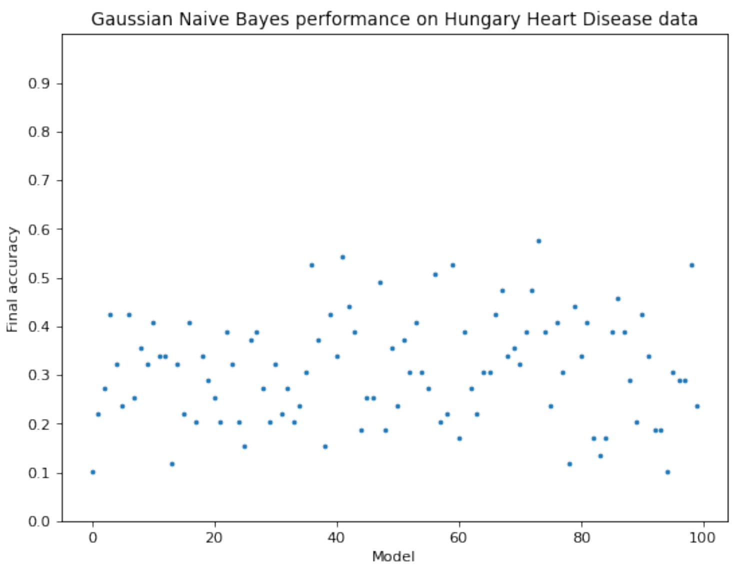


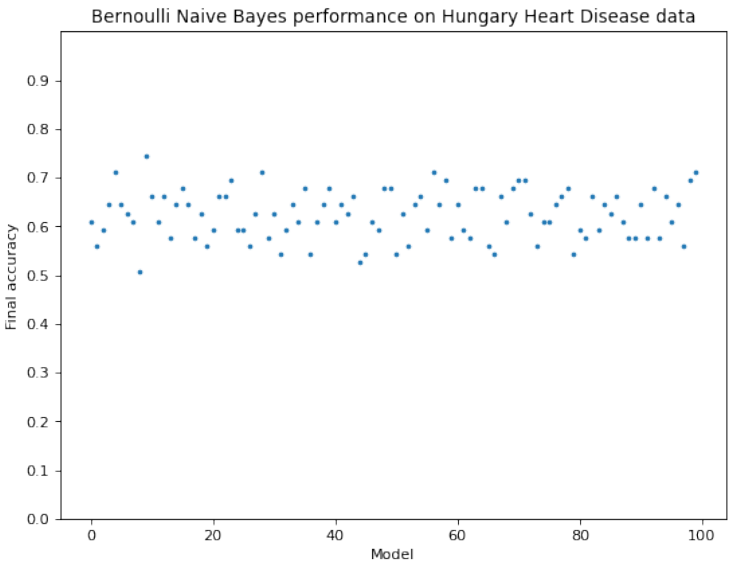


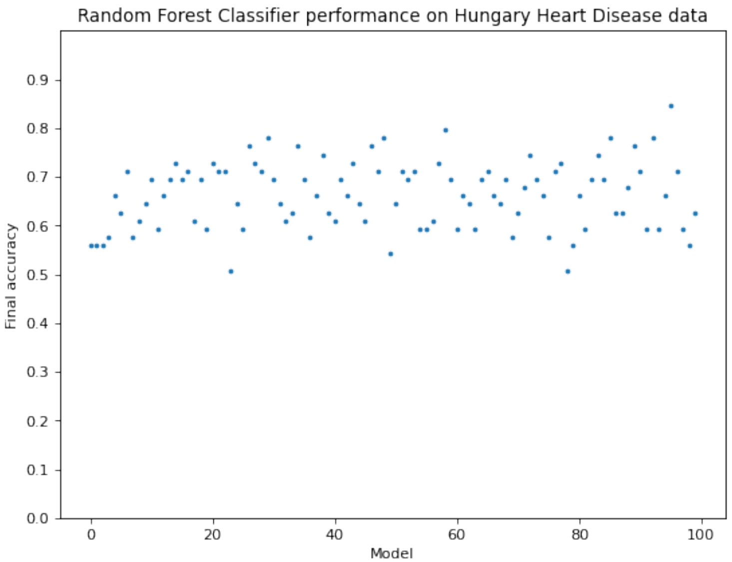


- 100 models each
- Best model accuracies
  - GNB: 36%
  - BNB: 56%
  - RF: 60%
- Relatively lower accuracies than on Cleveland Heart Disease
  - 273 missing values

Hungary Heart Disease

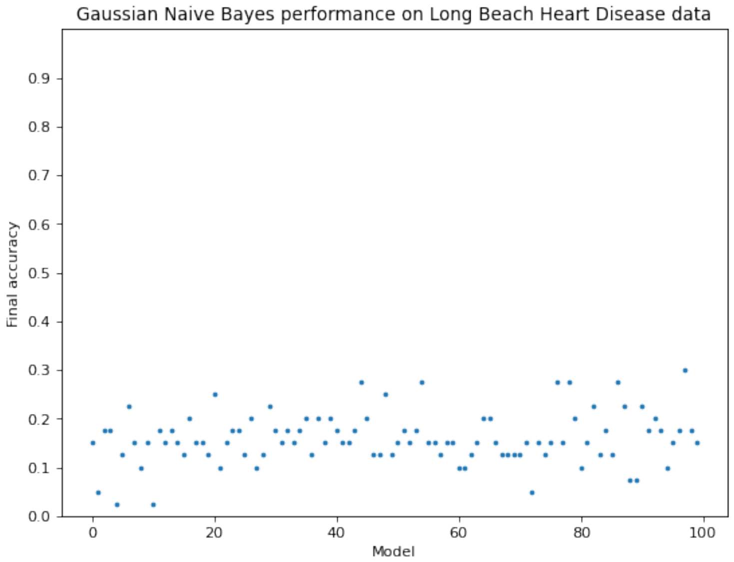


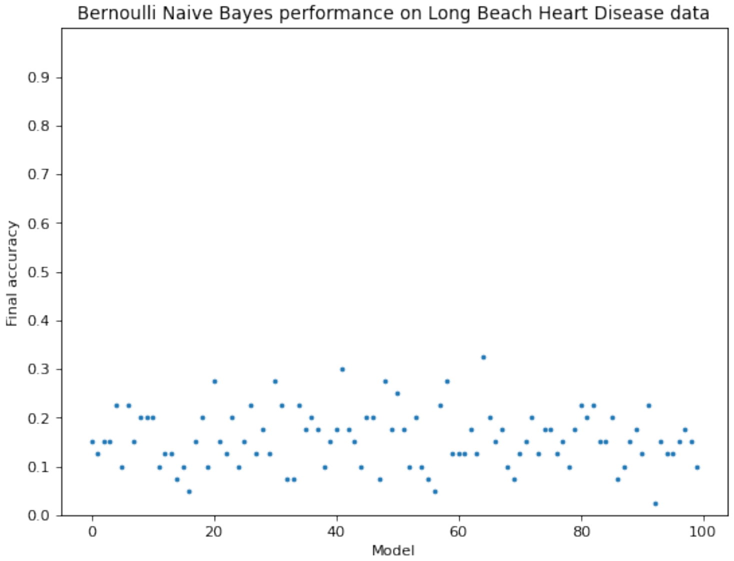


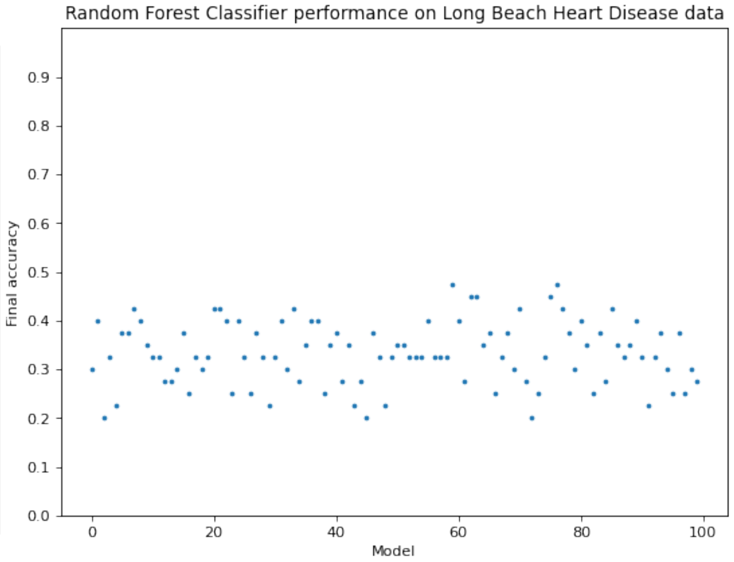


- 100 models each
- Best model accuracies
  - GNB: 57.63%
  - BNB: 74.58%
  - RF: 84.75%
- No missing values

Long Beach Heart Disease

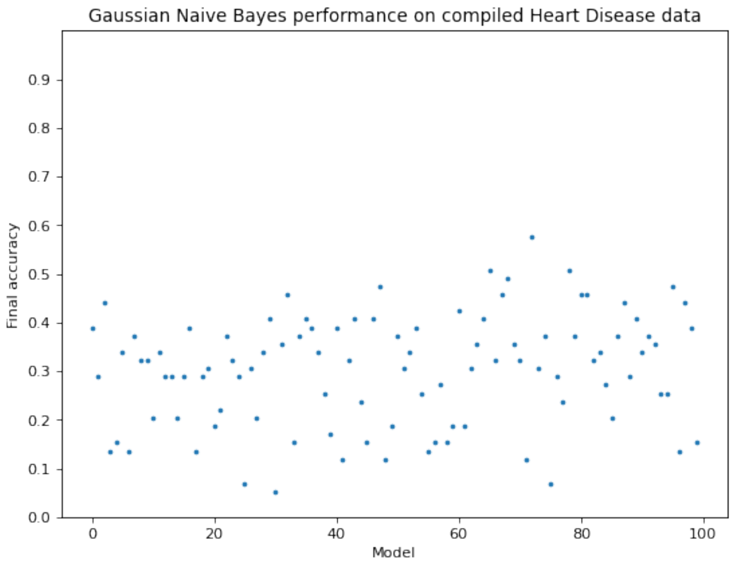


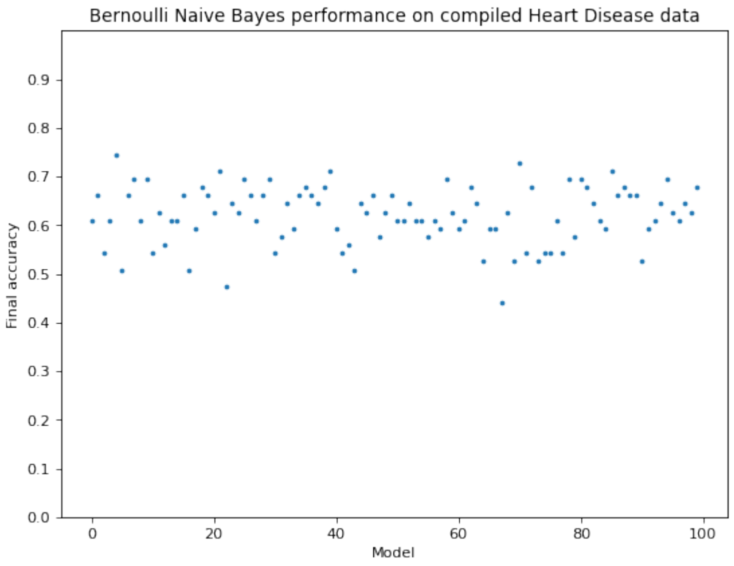


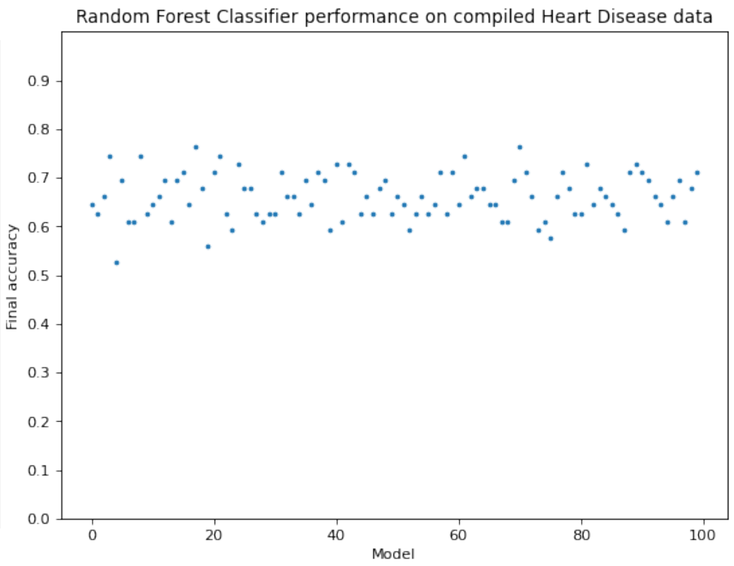


- 100 models each
- Best model accuracies
    - GNB: 30%
    - BNB: 32.5%
    - RF: 47.5%
- Lowest model accuracies (and most missing values) out of all Heart Disease subsets
    - 698 missing values

Compiled Heart Disease data







- 100 models each
- Best model accuracies
  - GNB: 57.62%
  - BNB: 74.58%
  - RF: 76.27%
- Compared to:
  - 65.57%, 67.21%, 72.13% (respectively) in our tests on Cleveland Heart Disease
  - 85%, 85%, 75% (respectively) in source publication

Conclusions

- **Recommendations for reproducible ML research in medicine**
  - Training and testing models on varied data sets
  - Evaluating problem-specific model reliability
  - Documenting detailed information: data preprocessing, model parameters, raw data and code
- **Need for defined reproducibility and transparency standards in ML for healthcare applications.**

Questions?

GitHub:

Email: hahmed@sandia.gov