An Analysis of Predictive Healthcare Model Transferability

Dayton Berezoski
Department of Computer Science and Engineering
Texas A&M University
College Station, TX
daytonbere@tamu.edu

I. Introduction

In the recent past, breakthroughs in the field of machine learning and artificial intelligence have helped pave the way for smarter and more efficient healthcare systems. Many of the leading research hospitals have been implementing these techniques, to do things such as provide a second opinion for physicians with predictive models, help patients understand their treatment with chatbots, and more [1]. However, the question can be posed regarding how smaller rural non-teaching hospitals that do not have access to these resources will be able to benefit from these advancements.

It is important to note that much of the progress in machine learning's applications in healthcare is mostly being put to use in only the institution that researched that topic, leaving a void for many rural hospitals. With an already documented disparity in the quality of rural versus urban healthcare systems, these advancements could prove to further this gap. This is proven by the fact that although 11% of the physician workforce practices in rural areas, 20% of Americans live in those same areas, and 100 rural hospitals have closed in the last ten years [2]. Furthermore, smaller rural non-teaching hospitals tend to see very different reasons why patients come to their facilities, posing unique problems for them that large urban teaching hospitals don't usually see. For example, pneumonia represents six percent of the hospitalizations that occur in smaller rural hospitals as opposed to the three percent of hospitalizations it makes up urban teaching hospitals [3]. Coupled with the fact that 51% of patients in small rural hospitals are over 65 years old against 37% in large urban teaching hospitals [3], underlines the fact that the two are facing very different healthcare problems.

The difference in urban and rural healthcare quality will continue to expand, as research will continue to be done on large urban teaching hospital settings while the rural healthcare systems will not have the resources to keep up.

II. Literature Review

A. Generalizability of Predictive Models for Intensive Care Unit Patients

This paper looks at training predictive models to predict ICU mortality and testing these same models with data gathered from other hospitals. The researchers use the eICU database to train a logistic regression model with data from a variety of hospitals across the United States and test the same models on held-out data. They were able to find that a model trained on this data was able to be transferred well to other hospitals. Overall motivating a need for data sharing across healthcare facilities [4].

B. An Empirical Framework for Domain Generalization in Clinical Settings

This paper uses developments in domain generalization to help boost the performance in clinical machine learning models. The data tested is over in-hospital mortality and chest X-rays features using a recurrent neural network. The data is taken from the eICU database for the in-hospital mortality and a combination of MIMIC-CXR, CheXpert, Chest-Xray8, and PadChest for the chest X-rays. They found that even with the developments in domain generalization, the models had a significant decrease in quality when tested over data it has not been trained on [5].

III. Problem Formulation

Model transferability is a key problem that if addressed, can begin to pave the way for rural hospitals to start catching up in machine learning advancements. The problem that this paper will be looking at is ICU mortality and what types of techniques and models will accurately predict this with an emphasis on performance on data from hospitals it had not been trained on.

The choice of looking at ICU mortality stems from the fact that it is a binary classification (0 - patient survives the ICU, 1 - patient dies in the ICU) and that it can utilize a few features to encompass a multitude of reasons why a patient might be in the ICU. For example, hypoxemia (low oxygen saturation) could be caused by a heart failure or lung conditions like asthma. This will allow the model to be able to make conclusions on the health of the patient from vitals, rather than having it to be told the diagnosis of the patient. This choice was also significant because it is a relatively basic thing to look for since ICU mortality happens in all hospitals and is something that is recorded in healthcare databases. This means that the focus can be put on how well the model performs when being used on data from different hospitals.

IV. Problem Solution

A. Overview

This experiment will consist of a train dataset from urban hospitals and a test dataset from a collection rural hospitals, being used across three different types of machine learning models to analyze how the models perform when transfered and provide a basis on model transferability for future studies. The models chosen for this paper is a logistic regression model, a gradient boosted model, and a vanilla neural network. The three models were chosen as they are relatively basic models as they will be able to encompass overall linear, tree-based, and neural network performance when transferred.

The models will be trained on the train dataset with a held-out validation set from the train dataset to evaluate the model's performance on the training data. The splits will be done in either a k-fold stratified coss-validation or a train-test split. This is to retrieve the hyper-parameters used to optimize the performance of the models on the training dataset, which simulates a well trained model on urban hospitals. Once the optimal hyper-parameters are found for a model, a new model is created with the same parameters to then be trained on the entirety of the train dataset. The model will be tested on the entirety of the test (rural) dataset, where the results will be measured against the initial model's performance on the validation dataset.

These procedures are done to ensure that the model is well trained on the training dataset to simulate a well-trained model from a large urban teaching hospital. This is so that any differences in model performance across datasets can be attested to the transferability of the model itself and not due to the model being unpredictable.

B. Model Performance Evaluation

The models will be scored on two different performance indicators being the Area-Under-the-Curve of a Receiver-Operating-Characteristic (AUC-ROC) and the Standard Mortality Rate (SMR). The idea of using the area under the curve of a precision recall curve came up, however when implementing it the standard functions used for calculating it were having compatibility issues with the predicted values from the neural network being compared against the actual values.

The AUC-ROC is a common performance metric used to measure the model's ability to separate the data into the binary labels. It is a scale from zero to one, where one means that the model is perfectly distinguishing the two classes, zero is model distinguishing the class with the opposite labels, and half is not able to distinguish between the two. The value is found by taking the true positive rate (TPR) and false positive rate (FPR), graphing TPR as a value of FPR and taking the area beneath the resulting curve.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{TN + FP} \tag{2}$$

(TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative)

The SMR value is a value that exists on the non-negative real-numbers, that evaluates the mortalities prediction by generating a ratio of the expected number of mortalities over the actual number. As such an SMR score close to one is good, above one predicts too many mortalities, and below one predicts too few.

$$SMR = \frac{ExpectedMortalties}{ActualMortalities} \tag{3}$$

Note: the SMR is only a ratio, it is not measuring the overall accuracy of the model. A model can be poorly trained producing many misclassifications, but still can have an SMR close to one.

The two metrics will work well together as the AUC-ROC will determine how overall separable the model is, while the SMR score will look specifically at the performance on data that results in mortality. The SMR is needed since the number of mortalities is much smaller than the number of surviving patients in both the train and test dataset, so this metric helps evaluate this smaller proportion of the data.

C. Logistic Regression Model

The logistic regression (LR) model was taken from sklearn.linear_model, version 0.24.2.

The model was first trained and tested on the train dataset, to evaluate how the model performs on data from the same hospital and retrieve optimal hyper-parameters. This was done by creating a 4-fold-stratified list to split the data with train and validate sets, and then for each fold train the model on the train data and evaluate it on the validation set. The k-fold-stratified function was taken from sklearn.model_selection. This is because LR is a simpler model allowing for quick training, which presented the opportunity of repeating it for an average value for the performance metrics.

Hyper-parameters:

- penalty: L2
- max_iter: 100000 - Chosen to make sure that the model could train under a large amount of data being fed in at once

D. Gradient Boosted Tree

The gradient boosted tree (GB) model was taken from sklearn.ensemble, version 0.24.2.

The model was first trained and tested on the train dataset, to evaluate how the model performs on data from the same hospital and retrieve optimal hyper-parameters. This was done using a train-test-split to split 20% of the train data into a validation set and use that to evaluate the model. The train-test-split function was taken from sklearn.model_selection. This was repeated three times to test performance of the max tree depth for values between three and five.

Note: SMR was not used in this evaluation since there would be times where very few actual mortalities were seen causing SMR to blow up in value.

Hyper-parameters:

- n_estimatiors: 1000
- learning_rate: 0.01
- max_depth: 3 - Not much change in validation performance occurred when this was changed, so 3 was chosen to help performance
- subsample: 0.5
- n_iter_no_change: 10

E. Neural Network

The neural network (NN) model was taken from pytorch, version 1.10.0.

The model was first trained on with a 20% train-test-split for training and validation dataset. The training function was run over 20 epochs, while comparing the accuracies and losses each time. This was done to retrieve the optimal hyper-parameters and the performance of the validation data.

Hyper-parameters:

- num_epochs: 20
- loss_function: Cross Entropy Loss - Taken from torch.nn
- learning_rate: 0.01
- optimization_function: Adam - Taken from torch.optim
- One hidden layer of 64 neurons
- Each layer in the NN uses a reLU activation function

V. Data Description

A. Survey of Data

Data is sampled from the MIMI-IV and eICU datasets to create the train and test data, respectively. This is done by hosting the two databases in the cloud using Google Big Query and using SQL commands to extract the necessary data. The features are a variety of commonly checked vitals, demographics, and responsiveness scores from the Glasgow Coma Scale [6]. These features were chosen as they are usually taken at multiple points in time for patients in the ICU.

The data is arranged so that each of the features (except for demographics) is recorded for the first and last time the patient is in the ICU, to get a sense of change over time when looking at the data. The features first recorded will have a "_f" appended to their name and the features last recorded will have a "_l" appended to their name. Only patients between 16 and 89 years old were considered for this study and each patient must have each of the features recorded for them to be used.

A full list of labels and meanings will be attached with Appendix A.

B. MIMIC-IV

Medical Information Mart for Intensive Care-IV (MIMIC-IV), is a database of deidentified data for patients admitted into the intensive care units at the Beth Israel Deaconess Medical Center [7]. Beth Israel Deaconess Medical Center is located in Boston, MA and is a large (673 bed) teaching hospital of Harvard Medical School [8]. Due to its affiliation with Harvard Medical School, Beth Israel Deaconess Medical Center has access to lots of advanced technology and resources to assist its physicians that other hospitals likely do not have access to. This makes MIMIC-IV the chosen training dataset, as it will serve as an extreme example of data from a large urban teaching hospital.

The database is split into three smaller databases called core, hosp, and icu. Core tracks the patient information taken during admissions and their exit. Hosp contains lab events for every patient in the hospital. Icu contains all of the vitals and other measurements for patients in the ICU. The data collected mostly joins the core and icu databases, resulting in 49503 total patients being used for the training dataset.

C. eICU

The eICU database is developed by Philips Healthcare contains deidentified data regarding patients admitted to different hospitals across the United States [9]. Since there are multiple hospitals included in the database, the first filter is to only allow data coming from hospitals with less than 500 beds or hopsitals that are non-teaching hospitals. This ensures that if there is a decrease in performance due to model transferability between large urban teaching hospitals and smaller rural non-teaching hospitals, it will easily be seen since the test data will not contain any data from similar healthcare facilities used in the training dataset.

The database is split into many smaller databases, however the ones most used are patient, physicalExam, hospital, vitalAperiodic, and vitalPeriodic. Joining these databases together allow for the necessary features to be extracted at different instances of time, resulting in 6497 patients being used for the testing dataset.

D. Preprocessing

First, careful inspection of the data has to be done in order to ensure that the units of measurement for features are the same between the two datasets. For example, the temperature from the MIMIC-IV database is recorded in Farenheit, while the temperature for the eICU database is recorded in Celsius.

The data is split into numeric and categorical sets temporarily to process the two seperately. Each of the numeric features for the MIMIC-IV and eICU were combined so they could all be normalized together to boost performance and decrease computational complexity. They were later separated back into their respective train and test datasets. The categorical data was one-hot encoded using the label encoder and one-hot encoder from sklearn.preprocessing to properly process them. It can be argued that the responsiveness (eye, verbal, motor) features from the Glasgow Coma Scale could be left as-is since there is an inherit numeric ordering to the values. However, the decision was made to encode them as the importance of the different values on the models is unknown to the user. The categorical and numeric is then rejoined resulting in 40 features for each patient of data.

The data has been preprocessed, however one additional step needs to be made for data going into the neural network. The data needs to be converted from a list into a tensor and then a data loader, however this is a relatively easy transformation that is does not need to be explained, but is worth noting.

VI. Results

The results here will be broken up by model and then into validation metrics, test metrics, and analysis of metrics. This is to show how the model behaves on test data from the hospital whose data it was trained on as compared to the test data from smaller rural non-teaching hospitals.

A. Logistic Regression

1) Validation Results:

- AUC-ROC: 0.8674
- SMR: 1.2197

2) Test Results:

- AUC-ROC: 0.6908
- SMR: 1.0464

3) Analysis: The validation results implies that the model performs decently well on data from the same hospital, with the only discrepancy being that the number of expected mortalities exceeds the actual number of mortalities by a fair margin, as indicated by the larger SMR. However, model has an improved SMR score on the test dataset with a lower AUC-ROC, implying that the model is having more trouble differentiating between the binary classification of mortality and the better SMR is just a result of luck with closer predicted to actual mortalities.

B. Gradient Boosted Tree

1) Validation Results:

- AUC-ROC: 0.9653
- SMR: Unable to be measured

2) Test Results:

- AUC-ROC: 0.7607
- SMR: 0.7782

3) Analysis: Due to a bug noted before, SMR was not able to be read in the validation results. However, with a high AUC-ROC of 0.9653, the validation SMR is likely close to 1. The test results imply that this model does much better at separating the binary classification of mortality than the logistic regression model. However, the difference is that SMR is much lower, meaning that this model is expecting more patients to survive the ICU than there actually are.

C. Neural Network

1) Validation Results:

- AUC-ROC: 0.8959
- SMR: 1.1093

2) Test Results:

- AUC-ROC: 0.7352
- SMR: 0.8866

3) Analysis: The neural network has a high AUC-ROC with an SMR slightly above 1 when measured over the validation results. This implies that the model is slightly more likely to predict that a dies when in actuality they do not. The test results have a lower AUC-ROC than the gradient boosted tree implying that the better SMR score is mostly due to misclassifications of both patients that do and do not die in the ICU.

D. Conclusions

In terms of performance, it seems that the gradient boosted model slightly outperforms the neural network which both outperform the logistic regression. This is seems to imply that the relationship between features and determining mortality is non-linear as the gradient boosted tree and neural network performed similarly and significantly better than the logistic regression model.

In both of the high performing models, the error came mostly from predicting that a patient would survive when that wasn't the case (from the low SMR values). This seems to imply that there is an underlying discrepancy between larger urban teaching hospitals and smaller regional non-teaching hospitals as according to the model it seems that the patient would have likely survived in the larger urban teaching hospital, likely due to the increase in resources.

VII. Limitations and Future Directions

A. Limitations

The largest limitation in this project is the large amount of features that appear in one database but not the other. For example, eICU records the mean arterial pressure (MAP) for all of the patients that appear in the dataset, however when the selected data from MIMIC-IV is filtered to only include patients with their MAP recorded the number of patients in the set goes from 49503 to around 6000. This means that this vital measurement is now unable to be used in the models. The same goes for glucose, FiO2, and pH, which are all in eICU, but did not appear for all patients in MIMIC-IV. Height and wieght were also not able to be found in the MIMIC-IV database, which would be crucial to the predictive model either because it is not recorded for many patients and that is why it was not able to be found or it is not recorded at all.

Another limitation is how the time series data is set up. Between MIMIC-IV and eICU, time series data is not recorded at the same frequency and for MIMIC-IV there is often times where one vital will be recorded, no other vitals are being recorded at the same time, as opposed to eICU where all periodic values are all recorded at the same time. While this doesn't rule out the possibility of using a recurrent neural network to make predictions it does complicate things enough for the possibility to be ruled out for this basic survey of models and their transferability.

B. Future Directions

An interesting next step is to create more advanced models, like a recurrent neural network mentioned before. The goal of this was to measure performance with basic models, so a natural next step would be to increase the complexity and see if these results still hold.

Another direction this could be taken is to filter eICU's data to create different test sets of data based on both hospital size and teaching status, rather than lumping all hospitals that aren't large urban teaching hospitals together like this project did. There might be some categories of hospitals that perform better than others by breaking down the problem further.

VIII. Conclusions

The main take away is that there is a start to making predictive healthcare models more transferable to smaller rural hospitals, since while there is a discrepancy in performance it seems to be consistent between the non-linear models and likely due to lack of hospital resources. While currently rural hospitals seems to be far behind in terms of machine learning resources as compared to larger urban teaching hospitals, this paper proves that models can be loosely transferred between the two. More research can be done in the future to analyze how these discrepancies can be mitigated, by adjusting the model when it's transferred.

References

[1] K. Sennaar, "How America's 5 Top Hospitals are using machine learning today," Emerj, 24-Mar-2020. [Online]. Available: https://emerj.com/ai-sector-overviews/top-5-hospitals-using-machine-learning/.

[2] R. Pifer, "Disparities between care in rural, urban areas getting worse," Healthcare Dive, 04-Dec-2019. [Online]. Available: https://www.healthcaredive.com/news/disparities-between-care-in-rural-urban-areas-getting-worse/568360/.

[3] M. J. Hall and M. Owings, "NCHS Data Brief, number 126, August 2013 - cdc.gov," Rural and Urban Hospitals' Role in Providing Inpatient Care, 2010, Apr-2014. [Online]. Available: https://www.cdc.gov/nchs/data/databriefs/db147.pdf.

[4] A. E. Johnson, T. J. Pollard, and T. Naumann, Generalizability of predictive models for intensive care unit patients.

[5] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi, An Empirical Framework for Domain Generalization in Clinical Settings.

[6] "Glasgow Coma Scale - Centers for Disease Control and ..." [Online]. Available: https://www.cdc.gov/masstrauma/resources/gcs.pdf.

[7] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-IV," MIMIC-IV v0.4, 13-Aug-2020. [Online]. Available: https://physionet.org/content/mimiciv/0.4/.

[8] "Beth Israel Deaconess Medical Center," BIDMC of Boston. [Online]. Available: https://www.bidmc.org/.

[9] T. Pollard, A. Johnson, J. Raffa, L. A. Celi, O. Badawi, and R. Mark, "EICU Collaborative Research Database," eICU Collaborative Research Database v2.0, 15-Apr-2019. [Online]. Available: https://physionet.org/content/eicu-crd/.

Appendix A
Data Labels

Label	Name	Type
age	Age	Numeric
gender	Gender	Categorical
bpm	Heart Rate	Numeric
resp_rate	Respiration Rate	Numeric
O2_sat	Oxygen Saturation	Numeric
tem	Temperature	Numeric
bps	Systolic Blood Pressure	Numeric
bpd	Diastolic Blood Pressure	Numeric
bpsm	Mean Blood Pressure	Numeric
eye_resp	Eye Responsiveness	Categorical
verbal_resp	Verbal Responsiveness	Categorical
motor_resp	Motor Responsiveness	Categorical

Note: A "_f" means first and "_l" means last recorded for time series data.