Clinical Predictive Modeling of Adverse Outcomes in Childbirth

By Krutika Pandit

Mentor: David Fenyo

Vilcek Institute NYU School of Medicine, New York, NY

A thesis in fulfillment of the Masters in Biomedical Informatics Degree

30 June 2022

Abstract

Predictive modeling of medical diagnoses, clinical risks, and adverse outcomes in healthcare is increasingly popular with the goal to reduce patient morbidity and mortality. One such risk during childbirth is high fetal birth weight, or macrosomia, known to be associated with complications such as cesarean section for failure to progress, postpartum hemorrhage, obstetric anal sphincter injuries (OASIS), and shoulder dystocia (SD). In this study, demographic and clinically relevant variables collected from a cohort of women aged 18 to 55 delivering at hospitals in the NYU Langone network from July 2013 to October 2018 (n = 31,580) were used to build and optimize a tree-based supervised machine learning model called XGBoost. The task included the binary classification of deliveries as having either a positive or negative diagnosis of macrosomia. Multiple techniques for data preprocessing and model building were explored. The best performing model with an F1 score of 0.734 was obtained with a balanced dataset, feature selection, and grid search optimization of five hyperparameters. After the search, the hyperparameters were set to the following values to produce the optimal results - maximum depth = 1, learning rate = 0.3, gamma = 0, lambda = 10, minimum child weight = 9, and number of estimators = 70. Maximum depth is the number of steps in the longest path between the root node and the leaf node. The learning rate determines the step size at each iteration while the model optimizes toward its objective. Gamma is the minimum loss reduction required to make a further partition on a leaf node of the tree. Lambda is the L2 regularization (ridge regression) term. Minimum child weight is the minimum sum of the instance weight (hessian) needed in a child. Lastly, the number of estimators is the number of trees in the ensemble and is equivalent to the number of boosting rounds.

I. Introduction:

1.1 Adverse Outcomes in Childbirth

There are a number of perinatal adverse outcomes including the presence of intrauterine fetal growth restriction (IUGR), low birth weight (LBW), preterm birth, stillbirth, and low Apgar score leading to an increased risk of neonatal intensive care unit (NICU) admissions.[1] With advances in medical technology, there has been a negative trend over the last few decades in the occurrence of some of these outcomes (e.g. stillbirth, preterm

birth rates); however, when present, infant morbidity and mortality are known to have a detrimental effect on the well-being of, both, the patient and physician.[2-5] The medical process of childbirth comes with uniquely high expectations of a perfect outcome and newborn. When combined with errors in disclosure, any outcome deviating from those expectations can result in litigation.[6]

Legal proceedings are commonplace in the high-risk practice of obstetrics, especially in developed countries.[7] Concerns within the medical community are particularly heightened due to poor legal knowledge among recent medical graduates.[8] One study noted that within a cohort of OB/GYN residents, 60% (n = 500) of respondents were unable to identify malpractice as a form of tort liability, 21% (n = 181) were unsure, and 44% (n = 378) reported receiving no medico-legal education during residency.[9] As dialogue continues about the implementation of a legal curriculum as part of post-graduate medical training, clinical decision support models using real-time data collection may prove beneficial to alleviate physician anxiety surrounding adverse patient outcomes.[10,11]

An assessment of delivery outcomes can begin by monitoring the risk of infant and maternal morbidity which is known to increase with birth weight.[12-14] A large cohort study of more than six million birth and infant death records demonstrated no difference in perinatal outcomes between the group weighing 4,000-4,500 g and the group weighing less than 4,000 g, but morbidity and mortality increased significantly in newborns weighing greater than 4,500 g.[15] Consequently, this study focuses on the risk factors associated with excessive fetal growth.

Fetal macrosomia is defined in absolute terms as a fetal weight of greater than 4,000 to 4,500 g, regardless of gestational age.[16] It can be subcategorized as symmetric or asymmetric macrosomia based on the proportionality in the growth of all fetal parameters. A variety of maternal factors predispose a newborn to both kinds of macrosomia, including constitutional factors, pre-existing diabetes, gestational diabetes mellitus (GDM), pre-pregnancy obesity, excessive gestational weight gain, abnormal fasting, postprandial glucose levels, dyslipidemia, a prior macrosomic newborn, and post-term pregnancy.[12,17-24] In the presence of these maternal factors, an accurate estimation of birth weight would prove beneficial in assessing the risk of adverse outcomes during delivery.

Prenatal estimation of fetal weight started in the 1960s, but was revolutionized when an algorithm for prediction was developed in 1984 based on the measurement of biparietal diameter, head circumference, abdominal circumference, and femur length.[25] Almost four decades later, the same algorithm is used as the basis for ultrasound imaging. The most accurate diagnosis of macrosomia, however, involves weighing the newborn after birth. Pre-partum imaging accuracy decreases with increasing fetal weight beyond 4,000 g.[26,27] For example, ultrasound-estimated fetal weight of more than 4,500 g accurately predicted birth weight of more than 4,500 g in only 33-44% of cases.[26-31] Magnetic resonance imaging (MRI) resulted in a lower relative error (2.6-3.7%) compared to two-dimensional ultrasound imaging (6.3-11.4%) when measurements were performed at less than one week from delivery, but challenges remain for this technique to be generalized for clinical implementation.[25]

Once diagnosed, macrosomia is associated with serious maternal risks including cesarean section for failure to progress, postpartum hemorrhage, obstetric anal sphincter injuries (OASIS), and shoulder dystocia (SD).[32,33] SD is defined as an obstruction of the descent of the anterior shoulder by the pubic symphysis. As a result of such an obstruction, fetal maneuvers are required and can lead to further complications. One study associated maternal symphyseal separation and transient lateral femoral cutaneous neuropathy with the aggressive hyperflexion of the maternal legs when attempting to rectify dystocia.[34] Two relatively recent studies also showed an increased risk of OASIS as a result of fetal manipulation.[35,36]

Macrosomia and SD not only introduce maternal risks but can also result in several fetal complications. The most commonly associated injuries include fracture of the clavicle and damage to the nerves of the brachial plexus, specifically at vertebrae C5 and C6.[37] This can lead to Erb-Duchenne palsy, transient paralysis of the arm that usually resolves over time.[38] In the presence of maternal diabetes, a birthweight of 4,500 g or more has been associated with a 20-50% chance of SD.[13,39] It is important to note that although macrosomia increases the risk of SD, most macrosomic newborns do not experience SD.[40] According to ACOG, most instances of SD occur unpredictably among newborns of normal birth weight.[41]

Other risks to the newborn include hypoglycemia, meconium aspiration, increased chance of admission to the NICU, and prolonged admission (more than three days) to the NICU.[12,42-44] Macrosomic newborns are also more likely to be overweight and obese later in life as compared to normal-weight newborns.[45,46]

Additionally, the duration of the SD alone is not a predictor of neonatal asphyxia or death.[47] In developed countries, the most prevalent risk factors associated with stillbirth are non-Hispanic black race, nulliparity, advanced maternal age, obesity, pre-existing diabetes, chronic hypertension, smoking, alcohol use, having a pregnancy using assisted reproductive technology, multiple gestation, male fetal sex, unmarried status, and past obstetric history.[48]

The interplay of maternal risks and infant complications associated with macrosomia makes it a prominent target for further investigation and a potential indicator of the need for medical intervention; however, the ineffective implementation of imaging to predict prenatal fetal birth weight poses a challenge in developing a plan for intrapartum management in patients with suspected macrosomia. Alternative non-invasive methods for a more accurate prenatal diagnosis are necessary.

1.2 Methods of Medical Intervention

There are several interventions for assistance with childbirth; however, this section reviews those specifically addressing macrosomia and SD. When the benefits of an expeditious delivery outweigh the risks of continuing the pregnancy, the use of labor induction is relatively common. More than 22% of all live births undergo this process, and the rate of induction in the U.S. has more than doubled since 1990.[49] As defined by ACOG, it is the use of medications or other methods to soften the cervix in preparation for labor. Literature surrounding its effect on the incidence of SD in term patients with suspected fetal macrosomia is inconsistent.[47] Some studies report an increased risk of cesarean delivery upon labor induction without a reduction in SD or newborn morbidity,[50-54] whereas others suggest a slight decrease or no effect on the risk of cesarean delivery.[55,56] Due to the lack of additional studies, ACOG continues to discourage induction of labor solely for suspected macrosomia at any gestational age.[47]

Most fetuses with macrosomia that are delivered vaginally do not experience SD.[47] Therefore, elective cesarean deliveries solely due to a prenatal diagnosis of macrosomia would disproportionately increase the rate of cesarean deliveries compared to the reduction in the rate of SD.[57-59] Although a pre-partum diagnosis of macrosomia may be inaccurate, cesarean birth reduces the risk of birth trauma and brachial plexus palsy. Therefore, it may be beneficial for newborns with suspected macrosomia who have an estimated fetal birth

weight of at least 5,000 g in women without diabetes and an estimated fetal birth weight of at least 4,500 g in women with diabetes.[60]

In observational studies, assisted vaginal deliveries of macrosomic fetuses using forceps or vacuum extraction in women with and without diabetes mellitus (DM) have shown to increase the risk of SD.[60] The odds of SD were found to be threefold to fivefold higher with vacuum extraction based on two population-based cohort studies as well.[61,62] The largest study and a meta-analysis found no increased risk of SD with forceps delivery.[61,63] Individual studies in the meta-analysis varied from a threefold increased risk with forceps use compared with a threefold decreased risk with vacuum use. ACOG recognises the lack of consistency in the literature evaluating the risk of complications with assisted vaginal birth based on EFW as well as the inherent inaccuracy in estimating fetal weight.[64] It recommends considering additional variables such as the adequacy of the maternal pelvis and the progress of labor, particularly during the second stage, for using operative methods in vaginal childbirth.

In the presence of macrosomia, prolonged first and second stages of labor are common.[60,65] This can lead to an increased risk of infection, chorioamnionitis, third-degree or fourth-degree perineal lacerations, and conversion to cesarean during labor.[13,32,65] Amniotomy, or the artificial rupture of membranes (AROM), may be used alone or in combination with oxytocin to treat slow labor progression.[66] A meta-analysis including 14 randomized trials showed a shortened duration of labor in association with an early AROM and early oxytocin policy.[67]

Episiotomy, or a small cut to widen the opening of the vagina, was routine in the management of SD, but with little scientific evidence.[68] Currently, ACOG advises the use of rotational maneuvers for SD management. It suggests an addition of nursing, obstetric, and anesthesia assistance professionals to the medical team and the cessation of applied fundal pressure while preparations are made for performing maneuvers. Several rotational maneuvers have been described.[47] However, according to ACOG, the McRoberts maneuver should be attempted first. In this maneuver, two assistants each grasp a maternal leg and sharply flex the thigh back against the abdomen causing cephalad rotation of the symphysis pubis and the flattening of the lumbar lordosis consequently freeing the impacted shoulder.[69,70]

Other maneuvers can be employed and repeated based on an assessment by the clinician.[47] With the Rubin maneuver, the healthcare provider places a hand in the vagina and on the back surface of the posterior fetal shoulder, then rotates it anteriorly towards the fetal face. With the Woods Screw maneuver, the health care provider instead rotates the fetus by exerting pressure on the anterior, clavicular surface of the posterior shoulder to turn the fetus until the anterior shoulder emerges from behind the maternal symphysis. Posterior axilla sling traction can also be applied using a size 12 or 14 French soft catheter.[71] The catheter is threaded to create a sling around the posterior shoulder, allowing the shoulder to be delivered by applying moderate traction to the sling. The Gaskin all-fours maneuver can be performed for women without anaesthesia.[72] It is important to note that no randomized controlled trials have compared maneuvers for SD alleviation, but studies reported improved communication, use of maneuvers, and documentation in SD handling as a result of obstetric simulations in medical educational institutions.[73-78]

Alternatively, a meta-analysis revealed that pre-pregnancy interventions such as exercise, low glycemic diet in women with GDM, and bariatric surgery in women with class two or class three obesity have shown to reduce macrosomia or large for gestational age (LGA) newborns without an increase in small for gestational age (SGA) or preterm delivery. [79,80] The same analysis also reported that prenatal exercise reduced the odds of cesarean delivery by 20%. Therefore, a pre-partum diagnosis of macrosomia would allow physicians to present relevant recommendations to the patient and ensure ample preparation for potential delivery maneuvers or procedures to subvert adverse birth outcomes.

1.3 Global Context and Confounders

After adjusting for population characteristics, the U.S. showed relatively high percentages of maternal BMI, cesarean deliveries, epidural anesthesia, and the use of other pharmacological relief compared to 13 high-income countries.[81] It showed relatively low percentages of low gestational age and birth weight. Data from the same study also showed relatively high rates of pre-labor cesarean delivery and low rates of assisted vaginal delivery in, both, nulliparous and multiparous women in the U.S.

Other confounding factors related to adverse outcomes in childbirth include adolescent pregnancies and the use of assisted reproductive technologies (ART). Both factors have been reported to increase maternal and perinatal complications necessitating advanced obstetric care.[82,83] Lastly, different racial groups showed

varied effects of GDM, pre-pregnancy obesity, and excessive pregnancy weight gain on an elevated risk of LGA infants.[84] To determine the best course of action in cases of suspected macrosomia, these variables may be considered in addition to the major factors outlined in sections 1.1 and 1.2.

1.4 Machine Learning Overview

Machine learning (ML), a subfield of artificial intelligence (AI), is rapidly changing the technological landscape.[85] Developments in fields such as natural language processing, computer vision, and automatic speech recognition (ASR) demonstrate that advances are not limited by specific data modalities. One study automated the identification of critical limb ischemia, a disease lacking a single definitive International Classification of Diseases (ICD-9 or ICD-10) code, using narrative clinical notes in the electronic health record (EHR).[86] Another study developed and validated a model for the detection of diabetic retinopathy using retinal fundus photographs.[87] Recurrent neural networks are able to focus on different locations of an input image and generate captions describing the image.[88] Finally, ASR can process human speech to facilitate documentation in multiple languages and assist individuals with voice, speech, or language disorders.[89]

The implementation of ML in healthcare has led to algorithms being used to assist medical decision-making. Large clinical datasets are now supported by the expansion of data gathering through wearable devices and extensive EHR systems in institutions offering health services.[11] For instance, a model was validated and integrated into the EHR in response to the COVID-19 pandemic to quantify the risk of adverse outcomes from the novel respiratory disease.[90] Preliminary results showed successful adoption into the clinical workflow using live patient data. Similarly, a model was able to preemptively identify adverse maternal outcomes in pre-eclampsia, a leading cause of maternal deaths.[91] Moreover, the great variety of sensors included in wearable devices (e.g. heart rate monitor, altimeter, GPS, gyroscope, microphone) facilitate the collection of a large set of variables.[92] With the availability of big data, decision support tools and user-friendly data visualization using ML techniques can be critical contributors to the multi-dimensional system of healthcare.[93]

ML can broadly be subdivided into three categories - supervised learning, unsupervised learning, and reinforcement learning.[94] Supervised learning algorithms, like the one implemented in this study, are task-driven and involve learning from a labeled training set to make predictions of discrete (classification) or

continuous (regression) variables. Some common algorithms in this category include decision trees, support vector machine (SVM), and neural networks. Unsupervised learning is applied when there is a need to understand the inherent structure of data without explicitly-provided labels.[95] Exploratory analysis (e.g. K-means clustering) and dimensionality reduction techniques (e.g. principal component analysis) are common use-cases of this type of learning. In reinforcement learning, an agent interacts with a virtual environment to determine the suitable action model to maximize total cumulative reward.[96] Its applications span across gaming, robotics, computer vision, and healthcare.

1.5 Clinical Prediction Models

ML models in a clinical setting aim to predict the future occurrence of specific outcomes for targeted medical intervention.[97] The need for previous patterns in data suggests the use of supervised ML in building such multivariate diagnostic and prognostic models.[98] As one of the simplest linear classification techniques, logistic regression is one such supervised model that has been used as a benchmark for sophisticated clinical trials.[99,100] It has been used by studies to predict instances of hemodynamic instability, imminent mortality, and the composite outcome of cardiac arrest, unplanned ICU admission, and mortality.[101-103] Like simple linear regression, logistic regression can use, both, discrete and continuous data points; however, its predictions are binary. The Wald's test is used to check if the explanatory variables in a logistic regression model are significant contributors for the prediction. Due to the binomial distribution along a logistic curve, least squares cannot be used to estimate the parameters. Instead, a technique called maximum likelihood estimation is applied.[104]

Least absolute shrinkage and selection operator (LASSO) regression is another ML technique that variably assigns priority to the factors affecting an outcome.[105] It completely excludes factors that are irrelevant to the outcome by reducing the associated parameters to a value of zero, thereby also reducing the variance in models. One study applied an adapted version of LASSO regression called priority-LASSO on an acute myeloid leukemia dataset consisting of clinical variables, cytogenetics, gene mutations, and expression variables.[106] The study structured its variables into blocks of different types (e.g. clinical, transcriptomic, methylation data) making the resulting predictions more practical and easy to interpret. The results showed a prediction accuracy similar to or even better than standard LASSO. Another study used LASSO in its ML

pipeline to develop an SD model using EHR factors such as maternal demographics, obstetric history, and sonographic evaluation within five weeks from delivery.[107] It was found to be more accurate than EFW alone in the prediction of SD or neonatal birth injury.

Neural networks, another form of supervised learning, may also be effective for clinical outcome prediction.[108] In summary, they are composed of interconnected layers (input, hidden, and output) of neurons that receive, process, and output a set of values. The layers are connected to each other by weights, and each layer is associated with a bias. Better predictions are made by the network when the parameters and hyperparameters are fine tuned based on the errors in prediction from previous iterations. Such networks have been adopted into clinical practice. For instance, one study used vital signs and laboratory values as predictors (features) of clinical deterioration events (ICU transfer and cardiac arrest) in patients hospitalized in the hematologic malignancy unit.[109] The resulting model outperformed the ViEWS model, an early warning scoring system that relies heavily upon vital sign abnormalities and assessment of mental status. The newly developed model and ViEWS model had positive predictive values of 82% and 24%, respectively. Another study reported improved predictions of long-term outcomes in ischemic stroke patients using neural networks.[110]

In decision trees, the type of model used in this study, data is continuously split based on a certain parameter to reach an outcome that can be discrete or continuous. Although tree-based methods have been criticized for overfitting and heavily relying on task-specific hand-engineered features, aggregating the predictions of an ensemble of trees using random forest (RF) or boosting has proved to be competitive against other supervised learning methods.[90,111] In RF based models, each tree is built using a random sample of features, and the feature space is split into more and smaller regions due to the lack of pruning. This combats overfitting to the training dataset while introducing diversity. RF based models have shown moderate predictive ability in postoperative outcomes of primary decompressive craniectomy, chemoradiotherapy outcomes, and long-term mortality and morbidity in stroke patients.[112-114]

Certain challenges are common while implementing the models described above. For instance, feature extraction is a key step in the ML pipeline and requires the use of domain knowledge, data standardization, and latent representation for optimally selecting relevant features.[90,115] In the case of classification, as in this study, class imbalance often needs to be addressed. The incidence of SD among vaginal births in the vertex

presentation was reported to range from 0.2-3%, a small percentage of the total population studied.[41,116] One study proposed a class-imbalance aware learning strategy called cross-coupling aggregation (Cocoa) to deal with this issue in multi-label learning.[117]

Additionally, with large datasets comes the problem of missing values. Often a result of inefficient communication following patient transfer between institutions, EHR incompleteness needs to be addressed before modeling.[118,119] Lastly, the accuracy of the recorded values must be assessed by checking agreement between different elements within the EHR (such as assigned diagnosis and supplied medications) or by verifying whether values are within expected ranges.[90,120] This study explores various methods for imputation, class imbalance, feature selection, and model optimization.

1.6 Performance Evaluation

The evaluation of classification-based algorithms can use a confusion matrix as its basis.[121] Classification metrics can also be calculated from true positives (TPs), false positives (FPs), false negatives (FNs), and true negatives (TNs). The use of each metric is determined by the purpose of the classifier. For example, if an experimental drug were to be tested, then a conservative test with few FPs would be required to avoid testing the drug on unaffected individuals.[122]

Four additional metrics for classification scenarios include accuracy, precision, sensitivity, specificity.[123,124] Accuracy is a ratio of the correctly predicted observations to the total number of observations. If the disease is rare, such as SD, predicting that all the subjects will not experience SD offers high accuracy but is not useful for diagnosis. As a result, it is not a good metric to use in case of class imbalance. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Sensitivity, or recall, is the TP rate whereas specificity is the TN rate. A receiver operating characteristic (ROC) curve plots the FP rate (1 - specificity) on the x coordinate and the TP rate (sensitivity) on the y coordinate. When plotted, the results from a random classifier would produce a linear graph with slope equals one. A good classifier produces an ROC curve bent towards the top-left corner of the graph demonstrating a low FP rate and a high TP rate. This results in a greater area under the ROC curve (AUC).[125] In case of high class imbalance, area under the precision recall curve (AUPRC) can also be measured to assess the predictive strength of the model.

F1 score is an additional metric for model performance evaluation.[126] In cases of class imbalance, the F1 score is a better metric than accuracy. It is defined as the harmonic mean of precision and recall, and it takes into account the different types of errors (FP or FN) made by the model. This study uses confusion matrices, AUROC curves, AUPRC curves, and F1 scores in its analysis.

1.7 Summary

Although uncommon, adverse outcomes during childbirth have lasting effects on the healthcare recipient and provider. High birth weight is a known risk factor for labor, and early detection of macrosomia and the chance of SD can provide the caregiver additional basis for preventative recommendations such as pre-pregnancy exercise. Since pre-partum ultrasound imaging has had limited success in detecting excessive fetal birth weight, ML models that can process, both, discrete and continuous clinical variables may be beneficial for decision support. Computational models may also directly be used to determine the best medical interventions during labor using all available clinical variables, not just EFW, while keeping in mind recommendations from ACOG and relevant literature.

This study used a boosted tree-based model (XGBoost) to predict the incidence of macrosomia using pre-partum EHR data. Ultimately, this project aims to incorporate the predictive modeling of various birth risks like macrosomia into the EHR for live detection.

II. Data:

Data was collected from women aged 18 to 55 delivering at hospitals in the NYU Langone network from July 2013 to October 2018 (n = 31,580). Several variables were recorded from the electronic health record (EHR) including demographic information, obstetric, medical, surgical, and family history, vital signs, laboratory results, labor medication exposures, and delivery outcomes. Macrosomia was defined in absolute terms as a fetal birth weight of greater than 4,000 to 4,500 g, regardless of gestational age. Based on

```
AGE_AT_DELIVERY
Gravida
Para
Term
Preterm
Abortions
Living_at_Delivery_Ct
RACE_White
RACE_African_American
RACE_Asian
RACE_Native_Hawaiian_or_Other_Pacific_Islander
RACE_Other_Race
ETHNICITY_Not_of_Spanish_Hispanic_Origin
ETHNICITY_Patient_Refused
ETHNICITY_Unknown
R_BMI
GA_IN_DAYS
Insulin
Insulin_NPH
Insulin_aspart
Insulin_detemir
Insulin_glargine
Insulin_lispro
Insulin_regular
Oral_glycemic
HEMOGLOBIN_THIRD_TRIMESTER
PMH_Macrosomia_or_LGA
Macrosomia
IUGR
IUFD_or_demise
Grand_multip
Apgar_1
Apgar_5
Baby_Sex
```

Figure 1. Final list of variables included in the study. All variables were standardized to float type.

domain knowledge, relevance, availability, and completeness of the EHR data, 34 features were selected for analysis. The cohort was limited to n = 498 to adjust for class imbalance with 248 macrosomic cases (~49.8%).

III. Methods:

Sixteen experiments were conducted and stratified as shown in figure 2. Across layer one (L1), model performance using the fancyimpute (FI) tool was compared against performance using the in-built XGBoost "missing" parameter (XGB MP) for dealing with NaN values in the data.[127,128] Across layer two (L2), experiments were conducted with and without class imbalance (CI) further illustrated in figure 3. In cases with CI (experiments 5-8 and 13-16), the XGBoost scale_pos_weight hyperparameter was set to 126, the inverse of the class distribution; in the absence of CI (experiments 1-4 and 9-12), scale_pos_weight was set to 1. Addressing CI during initial preprocessing by only including an equal number of samples from each class dramatically reduced the sample size from n = 31,580 to n = 498. Across layer three (L3), experiments were conducted with and without feature selection (FS). Scikit-learn's SelectFromModel() tool was used for this step.[129] Across layer four (L4), grid search (GS) and Bayesian (Bay.) optimization techniques were compared. Scikit-learn's GridSearchCV() tool and a BayesianOptimization object were used for this step.

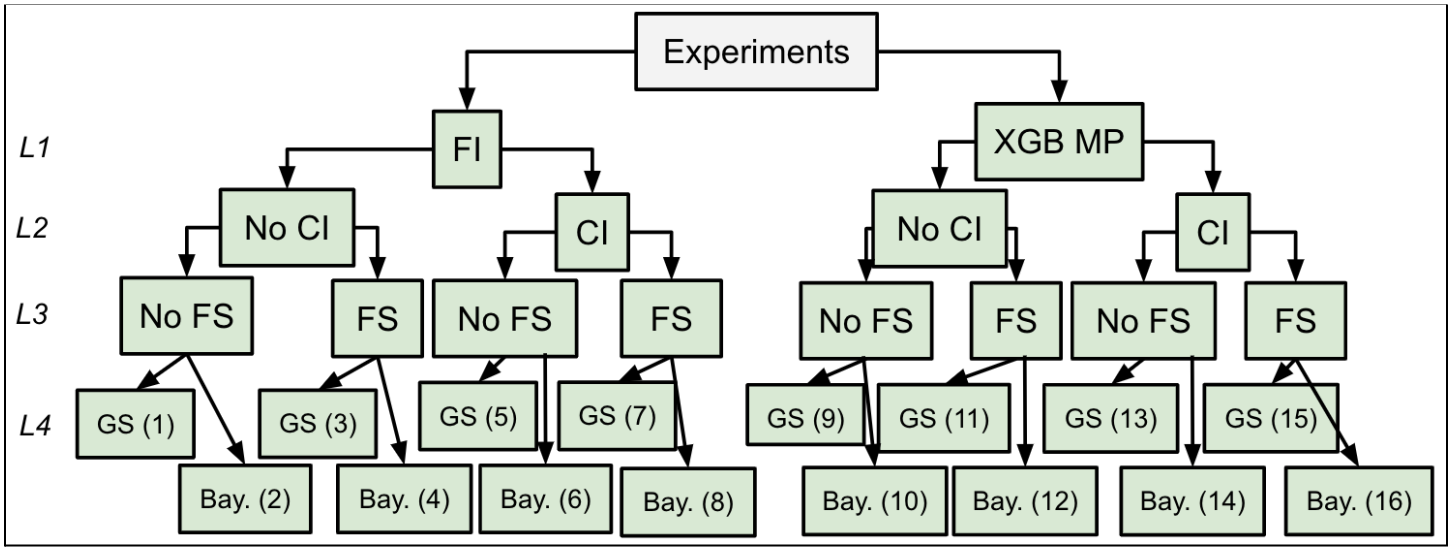

Figure 2. Experimental Design. Performance from sixteen experiments was compared using a combination of various imputation, weighted loss, feature selection, and optimization techniques to obtain the best results. In L4, the numerical values correspond to each of the sixteen experiments conducted.

3.1 Data Preprocessing

JupyterLab was used with the following packages for data cleaning and preprocessing - NumPy, pandas, regex, and fancyimpute. Variables with greater than 50% missing data and redundancy were excluded. One hot encoding was used to allocate a separate column for unique values in categorical variables. Data types were standardized within columns, and spaces in strings within the data frame and column names were converted to underscores. Missing data were either imputed using fancyimpute which regressively models each feature with missing values as a function of other features in a round-robin fashion, or handled using the in-built “missing” XGBoost parameter.

3.2 Model Building

Supervised machine learning was implemented to create a classification model based on training validation, and testing cohorts split using a 60-20-20 ratio. Samples were stratified based on the macrosomia label. Figure 3A and 3B show balanced and imbalanced classes between all three cohorts, respectively.

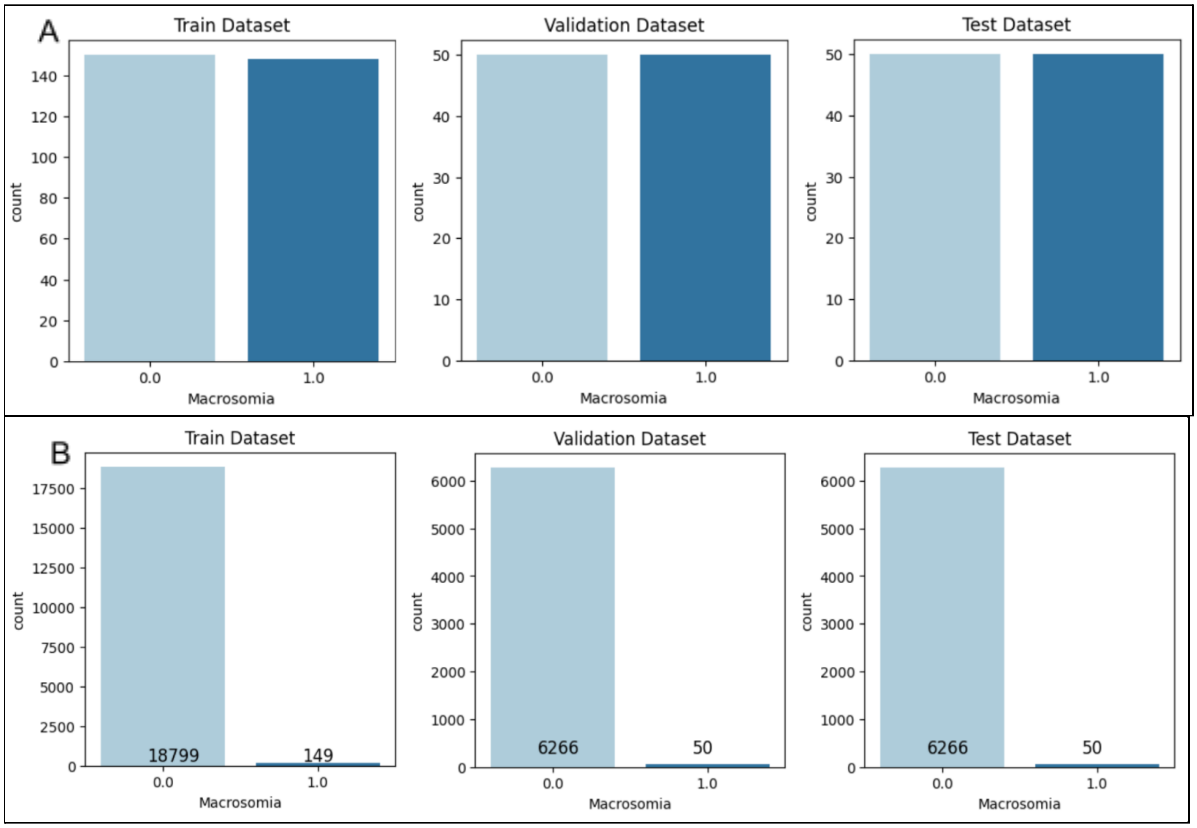

Figure 3. Value counts for balanced (A) and imbalanced datasets (B). Samples were stratified by binary prediction label, incidence of macrosomia.

Scikit-learn and XGBoost libraries were used to build the model.[128,129] Initially, default parameters were used to fit the XGBClassifier to the train dataset with early stopping after ten epochs. The objective parameter was set to 'binary:logistic' with 'seed = 42'. The model was evaluated on the validation set.

3.3 Optimization

Next, the GridSearchCV package from Scikit-learn was used to compute optimum values for the hyperparameters in experiments 1, 3, 5, 7, 9, 11, 13, and 15. A list of potential values for the following hyperparameters was first passed through the grid search with five-fold cross validation - maximum depth, learning rate, gamma, lambda, minimum child weight, and number of estimators. Maximum depth is the number of steps in the longest path between the root node and the leaf node. For example, in figure 4A, the depth is 5. Deeper trees can increase performance by adding complexity, but can also increase the chance of overfitting. The learning rate determines the step size at each iteration while the model optimizes toward its objective. A low learning rate makes computation slower, and requires more rounds to achieve the same reduction in residual error as a model with a high learning rate; however, it optimizes the chances to reach the best optimum.[130] Gamma is the minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be. Lambda is the L2 regularization (ridge regression) term. It can be used when training data is limited. Minimum child weight is the minimum sum of the instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than the set value for this hyperparameter, then the building process will give up further partitioning. Lastly, the number of estimators is the number of trees in the ensemble and is equivalent to the number of boosting rounds.[131]

The values for each hyperparameter were evaluated against the validation set to output a single optimal value. If a value from either extreme of the range passed was reported to be optimal, the range was readjusted to include higher/lower values for the next round of validation. If a central value was determined to be optimal, the value was kept constant for the subsequent round. Once the optimal values for each hyperparameter were determined, the model was fit to the train dataset again and evaluated on all three cohorts to check for regularization.

In experiments 2, 4, 6, 8, 10, 12, 14, and 16, the bayes_opt package was installed to create a BayesOptimization object and search the hyperparameter space.[132] The same five hyperparameters listed above were explored within the following ranges - maximum depth: (3, 10), learning rate: (0, 1), gamma: (0, 5), lambda: (3, 20), minimum child weight: (1, 10), and number of estimators: (50, 350).

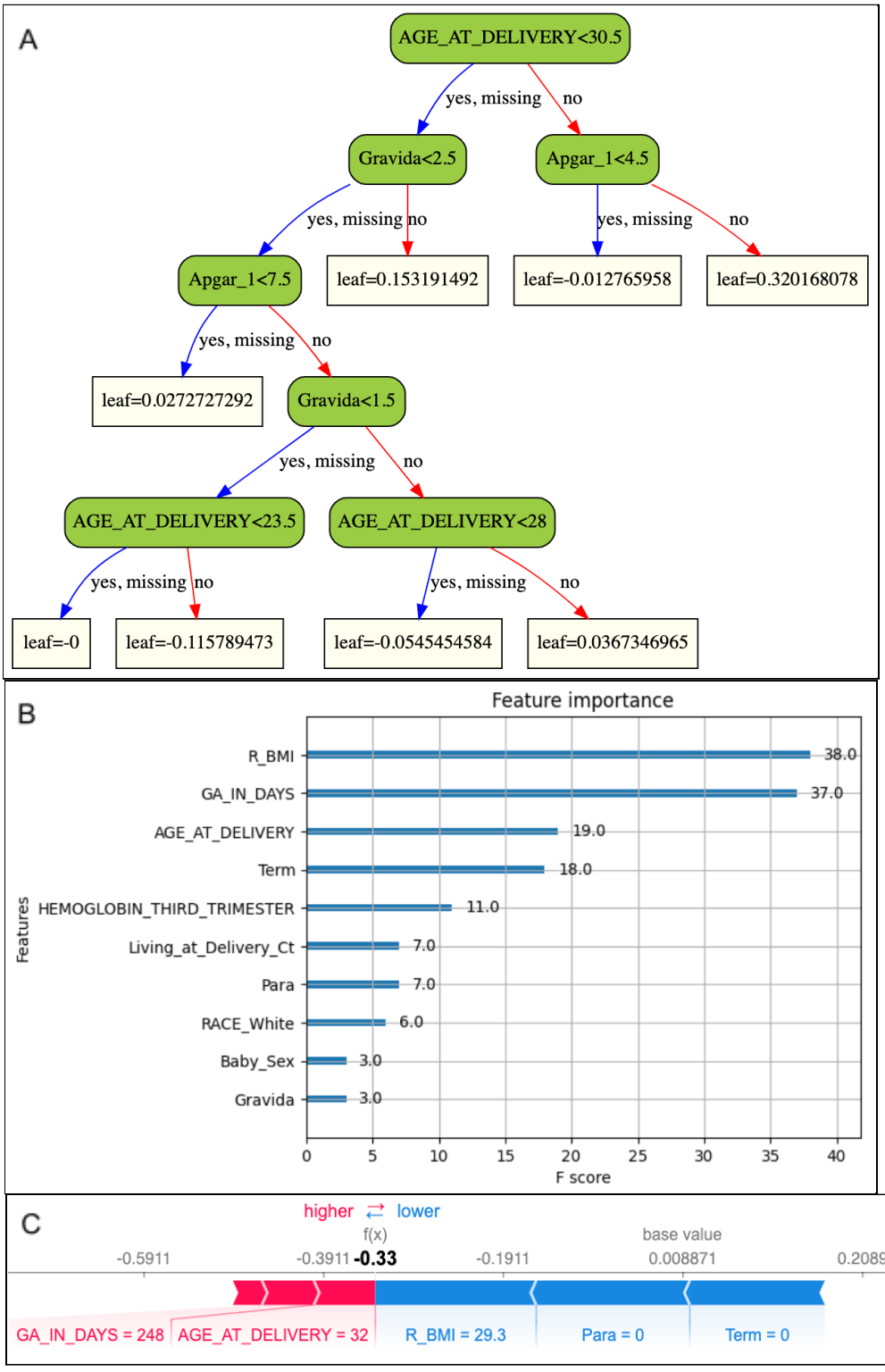

Figure 4. Methods for interpretability. (A) Example decision tree. Root nodes and leaf nodes are shown in green and beige, respectively. **(B)** Feature importance barplot example from experiment 12. Features with higher predictive value correspond to a high F score. **(C)** SHAP values example from experiment 1. Feature values contributing to a push towards a positive prediction are in pink, and their size shows the magnitude of the feature's effect. Feature values contributing to a push towards a negative prediction are in blue. The biggest predictive impact comes from an "R_BMI" value of 29.3; however, the "AGE_AT_DELIVERY" value of 32 also has a substantial effect in the opposite direction.

3.4 Interpretability

Example decision trees from all XGBoost models were visualized using the package Graphviz4 to translate the model methodology.[133] As shown in figure 4A, each node includes a threshold splitting the input into two subsequent groups to create leaves containing the most homogenous samples. The incremental probabilities from each leaf are added to the probabilities from all the other trees to provide a final probability that an observation is of one class or the other.

Since decision trees produced by the XGBoost model do not provide an adequate breakdown of the factors contributing to the predictions, feature importance bar plots and SHapley Additive exPlanations (SHAP) values were visualized.[134,135] Altogether, they can be incorporated into EHR systems such as Epic for live interpretation. Figures 4B and 4C show examples of these visuals from experiments 1 and 2, respectively.

IV. Results:

After evaluating the model performance for all sixteen experiments outlined in figure 2, experiment 11 produced the highest AUROC and F1 score. Missing data was imputed with zeros and the XGBoost "missing" parameter was used to identify missing values while training. Feature selection with grid search was implemented. Figure 5 shows the AUROC and AUPRC curves for experiment 11 before and after optimization.

Optimization using grid search decreased overfitting to the train dataset as demonstrated by similar AUROC, AUPRC, and F1 scores between the train, validation, and test sets. Overall, there was a 0.155 increase in AUROC and 0.116 increase in AUPRC. F1 score increased by 0.154. The evaluation metrics are further summarized in Table 1.

	AUROC		**AUPRC**		**F1 score**	
	Before	After	Before	After	Before	After
Train	**1.000**	0.774	**1.000**	0.777	**1.000**	0.728
Valid	0.705	**0.723**	0.675	**0.766**	0.602	**0.673**
Test	0.657	**0.812**	0.688	**0.804**	0.580	**0.734**

Table 1. Evaluation Metrics Summary. AUROC, AUPRC, and F1 score were recorded before and after optimization of XGBoost hyperparameters using grid search. The higher values between before and after are bolded.

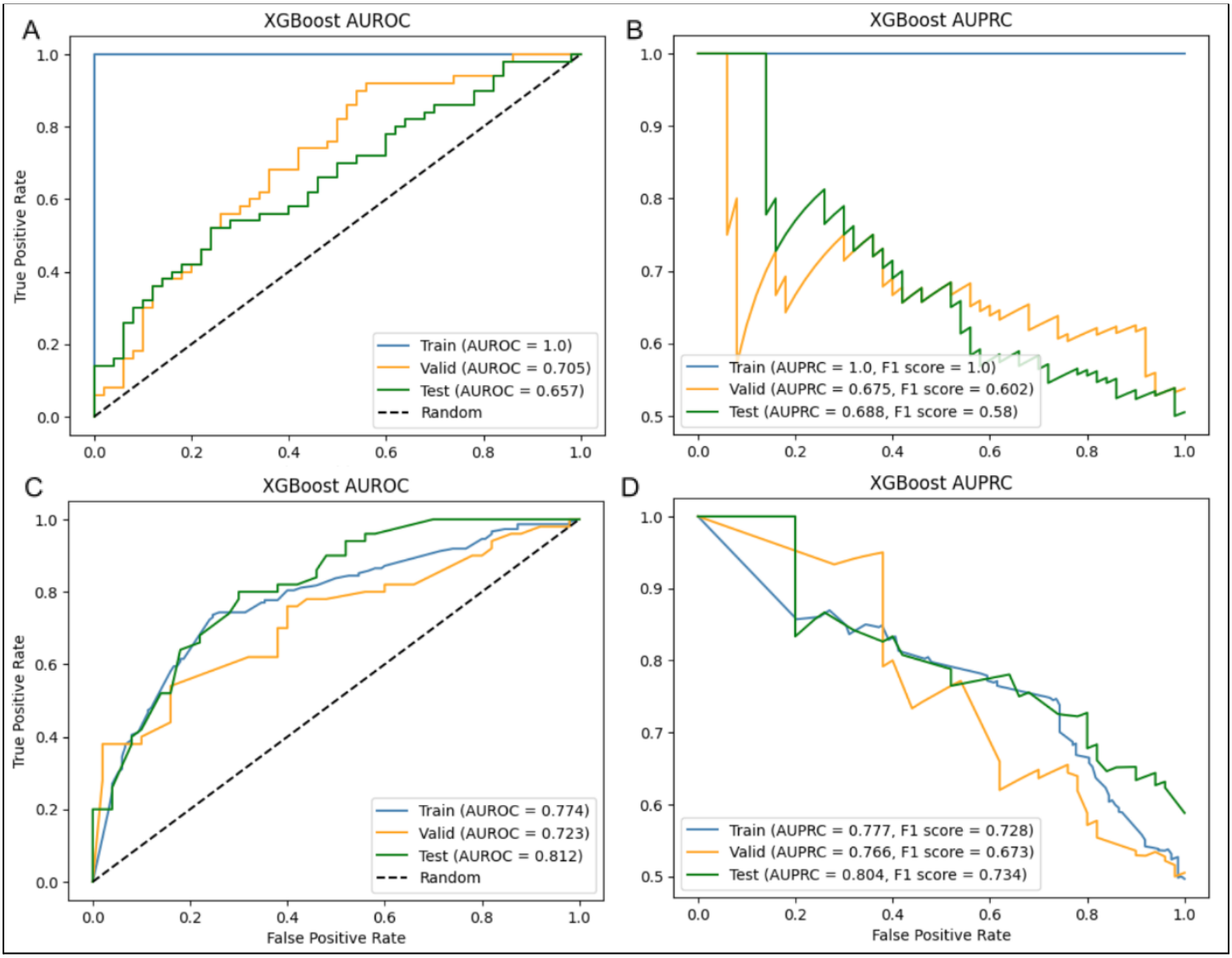

Figure 5. AUROC and AUPRC plots before and after optimization in experiment 11. A trained XGBoost classification model was used to predict the incidence of macrosomia. The false positive rate (fpr) and true positive rate (tpr) were calculated using the roc_curve() method. AUROC score was calculated using the roc_auc_score() method. The precision and recall were calculated using the precision_recall_curve() method. AUPRC score was calculated using the auc() method. F1 score was calculated using the f1_score() method. All methods were imported from the sklearn.metrics module. **A** and **B** show plots before optimization; **C** and **D** show plots after optimization of hyperparameters.

Shown in figure 6, true and predicted labels for the incidence of macrosomia were also plotted as confusion matrices before and after optimization for experiment 11. Optimization using grid search increased agreement between true and predicted labels for both classes by 0.04 for the “No Macrosomia” class and by 0.22 for the “Macrosomia” class.

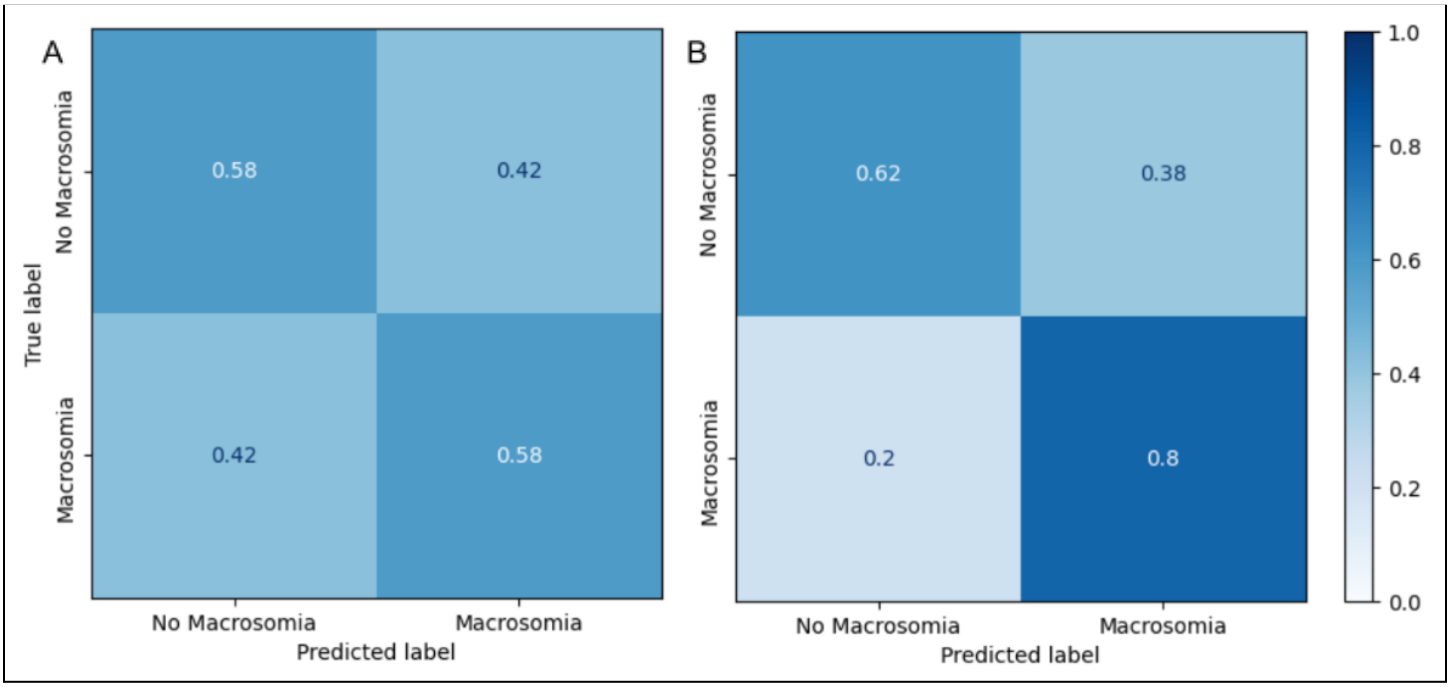

Figure 6. Confusion matrices before (A) and after (B) optimization. Count values displayed by class were standardized to a range of zero and one. Darker color represents greater agreement between true and predicted labels.

V. Discussion:

Using pre-partum features listed in figure 1, an XGBoost algorithm was able to predict the incidence of macrosomia with an F1 score of 0.734. The best evaluation metrics were obtained with a balanced dataset, feature selection, and grid search optimization of five hyperparameters. After the search, the hyperparameters were set to the following values to produce the optimal results - maximum depth = 1, learning rate = 0.3, gamma = 0, lambda = 10, minimum child weight = 9, and number of estimators = 70.

One study created a similar optimized XGBoost model to predict a heart disease diagnosis.[136] The study was able to obtain an F1 score of 0.9056 with Bayesian optimization. It included similar preprocessing of 14 numerical and categorical variables; however, twice as many hyperparameters were tuned in comparison to the ones in this study. Categorical variables were processed using one-hot encoding, and the entire dataset was fairly complete with only six missing values. In the heart disease study, 20-fold cross validation was performed compared to 5-fold cross validation used in this analysis.

Although this study included a large pool of deliveries from the NYU Tisch hospital (n = 31,580), the completeness of data heavily guides model performance. With large amounts of missing data, the predictive ability of the selected features may be reduced. Furthermore, NYU Tisch hospital oversees healthcare for patients from a relatively affluent network. Socio-economic status is known to be related to health outcomes.[137] Therefore, even with a highly efficient model, the parameters would need to be readjust to be applied to any population of births outside the studied community.

Furthermore, the model in this study was limited to 34 demographic and clinical pre-partum variables (including one-hot ended variables). If possible, future work may be able to utilize additional data types such as imaging and time series for better performance.[138] Imaging data is collected throughout the course of pregnancy with recorded values for biparietal diameter, head circumference, abdominal circumference, and femur length. The images themselves may also be inputted into convolutional neural networks (CNNs) known for their ability to handle high dimensional data to make predictions. Variables recorded as time series including temperature, heart rate, blood pressure, and pulse oximetry may also provide additional context to the health of the patient and, consequently, the weight of the fetus.

Macrosomia, or high fetal birth weight, is only one of the many risk factors during delivery. Techniques explored in this study can be transferred to other outcomes as well to provide a comprehensive live view of the patient's health status. Once incorporated into an EHR system like Epic, cluster randomized trials (CRTs) can be conducted to statistically analyze the effect of model integration on delivery outcomes. Random sampling to create control and test clusters can include stratification of the physician population to allow the control group outcomes to closely represent the counterfactual outcomes when calculating the average effect of the intervention. Randomization would be done on a physician level instead of an individual visit level to minimize contamination in the case of patients visiting the same physician on multiple occasions. Each physician will either receive all hidden or all not hidden predictions. Control delivery outcomes serving as the counterfactuals can be compared to the delivery outcomes with model intervention using Welch's t-test. This version of the independent group t-test takes into account the remaining inter-cluster variances between the two groups and adjusts the p-value accordingly.

Overall, the results from this study can be extended by several means, including greater completeness of data, including additional types of data such as imaging and time-series, tuning more hyperparameters, using 20-fold cross validation, and exploring a more diverse patient population. The techniques applied in this analysis can be expanded to other adverse outcomes in childbirth as well. Once optimized, model performance can be made interpretable with packages such as SHAP and incorporated into live patient care to guide physician recommendations throughout pregnancy and surgical preparations during delivery.

Acknowledgements

Thank you to Dr. David Fenyo and Dr. Wenke Liu for the continual mentorship and support throughout this thesis project. Thank you to Dr. Myah Griffin for help with identifying variables from the electronic health record relevant to this study.

References:

[1] Getaneh, T., Asres, A., Hiyaru, T. & Lake, S. Adverse perinatal outcomes and its associated factors among adult and advanced maternal age pregnancy in Northwest Ethiopia. Sci Rep 11, 14072 (2021).
CME Info - Child Mortality Estimates. https://childmortality.org/data/United%20States%20of%20America.
[2] CME Info - Child Mortality Estimates. https://childmortality.org/data/United%20States%20of%20America.
[3] Hamilton, B., Martin, J. & Osterman, M. Births: Provisional Data for 2020. https://stacks.cdc.gov/view/cdc/104993 (2021) doi:10.15620/cdc:104993.
[4] Goldberg, R. M., Kuhn, G., Andrew, L. B. & Thomas, H. A. Coping with medical mistakes and errors in judgment. Ann Emerg Med 39, 287–292 (2002).
[5] Disclosure and Discussion of Adverse Events | ACOG. https://www.acog.org/en/clinical/clinical-guidance/committee-opinion/articles/2016/12/disclosure-and-discussion-of-adverse-events.
[6] Chou, M. M. Litigation in obstetrics: a lesson learnt and a lesson to share. Taiwan J Obstet Gynecol 45, 1–9 (2006).
[7] Adinma, J. Litigations and the Obstetrician in Clinical Practice. Ann Med Health Sci Res 6, 74–79 (2016).
[8] Blanchard, M. H. et al. Impact of the medical liability crisis on postresidency training and practice decisions in obstetrics-gynecology. J Grad Med Educ 4, 190–195 (2012).
[9] Mathew, S., Samant, N., Cooksey, C. & Ramm, O. Knowledge, Attitudes, and Perceptions About Medicolegal Education: A Survey of OB/GYN Residents. Perm J 24, 19.217 (2020).
[10] Evans, A. & Refrow-Rutala, D. Medico-legal education: a pilot curriculum to fill the identified knowledge gap. J Grad Med Educ 2, 595–599 (2010).
[11] Razavian, N. et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. npj Digit. Med. 3, 1–13 (2020).
[12] The Apgar Score | ACOG. https://www.acog.org/en/clinical/clinical-guidance/committee-opinion/articles/2015/10/the-apgar-score.
[13] Nesbitt, T. S., Gilbert, W. M. & Herrchen, B. Shoulder dystocia and associated risk factors with macrosomic infants born in California. Am J Obstet Gynecol 179, 476–480 (1998).
[14] Doty, M. S., Chen, H.-Y., Sibai, B. M. & Chauhan, S. P. Maternal and Neonatal Morbidity Associated With Early Term Delivery of Large-for-Gestational-Age But Nonmacrosomic Neonates. Obstet Gynecol 133, 1160–1166 (2019).
[15] Zhang, X., Decker, A., Platt, R. W. & Kramer, M. S. How big is too big? The perinatal consequences of fetal macrosomia. Am J Obstet Gynecol 198, 517.e1–6 (2008).
[16] Macrosomia - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/medicine-and-dentistry/macrosomia.
[17] Ehrenberg, H. M., Mercer, B. M. & Catalano, P. M. The influence of obesity and diabetes on the prevalence of macrosomia. Am J Obstet Gynecol 191, 964–968 (2004).
[18] Ferraro, Z. M. et al. Excessive gestational weight gain predicts large for gestational age neonates independent of maternal body mass index. J Matern Fetal Neonatal Med 25, 538–542 (2012).
[19] Alberico, S. et al. The role of gestational diabetes, pre-pregnancy body mass index and gestational weight gain on the risk of newborn macrosomia: results from a prospective multicentre study. BMC Pregnancy Childbirth 14, 23 (2014).
[20] Goldstein, R. F. et al. Association of Gestational Weight Gain With Maternal and Infant Outcomes: A Systematic Review and Meta-analysis. JAMA 317, 2207–2225 (2017).
[21] Farrar, D. et al. Hyperglycaemia and risk of adverse perinatal outcomes: systematic review and meta-analysis. BMJ 354, i4694 (2016).
[22] Wang, J. et al. Gestational dyslipidaemia and adverse birthweight outcomes: a systematic review and meta-analysis. Obes Rev 19, 1256–1268 (2018).
[23] Marshall, N. E. et al. The Association between Maternal Height, Body Mass Index, and Perinatal Outcomes. Am J Perinatol 36, 632–640 (2019).
[24] Santos, S. et al. Impact of maternal body mass index and gestational weight gain on pregnancy complications: an individual participant data meta-analysis of European, North American and Australian cohorts. BJOG 126, 984–995 (2019).

[25] Kadji, C. et al. Magnetic resonance imaging for prenatal estimation of birthweight in pregnancy: review of available data, techniques, and future perspectives. Am J Obstet Gynecol 220, 428–439 (2019).
[26] Scioscia, M., Vimercati, A., Ceci, O., Vicino, M. & Selvaggi, L. E. Estimation of birth weight by two-dimensional ultrasonography: a critical appraisal of its accuracy. Obstet Gynecol 111, 57–65 (2008).
[27] Zafman, K. B., Bergh, E. & Fox, N. S. Accuracy of sonographic estimated fetal weight in suspected macrosomia: the likelihood of overestimating and underestimating the true birthweight. J Matern Fetal Neonatal Med 33, 967–972 (2020).
[28] Scioscia, M., Vimercati, A., Ceci, O., Vicino, M. & Selvaggi, L. E. Estimation of birth weight by two-dimensional ultrasonography: a critical appraisal of its accuracy. Obstet Gynecol 111, 57–65 (2008).
[29] Sandmire, H. F. Whither ultrasonic prediction of fetal macrosomia? Obstet Gynecol 82, 860–862 (1993).
[30] Chauhan, S. P. et al. Limitations of clinical and sonographic estimates of birth weight: experience with 1034 parturients. Obstet Gynecol 91, 72–77 (1998).
[31] Aviram, A. et al. Different formulas, different thresholds and different performance-the prediction of macrosomia by ultrasound. J Perinatol 37, 1285–1291 (2017).
[32] Rossi, A. C., Mullin, P. & Prefumo, F. Prevention, management, and outcomes of macrosomia: a systematic review of literature and meta-analysis. Obstet Gynecol Surv 68, 702–709 (2013).
[33] Beta, J. et al. Maternal and neonatal complications of fetal macrosomia: cohort study. Ultrasound Obstet Gynecol 54, 319–325 (2019).
[34] Gherman, R. B., Ouzounian, J. G., Incerpi, M. H. & Goodwin, T. M. Symphyseal separation and transient femoral neuropathy associated with the McRoberts' maneuver. American Journal of Obstetrics and Gynecology 178, 609–610 (1998).
[35] Gachon, B., Desseauve, D., Fritel, X. & Pierre, F. Is fetal manipulation during shoulder dystocia management associated with severe maternal and neonatal morbidities? Arch Gynecol Obstet 294, 505–509 (2016).
[36] Gauthaman, N., Walters, S., Tribe, I.-A., Goldsmith, L. & Doumouchtsis, S. K. Shoulder dystocia and associated manoeuvres as risk factors for perineal trauma. Int Urogynecol J 27, 571–577 (2016).
[37] Ahn, E. S. et al. Neonatal clavicular fracture: recent 10 year study. Pediatr Int 57, 60–63 (2015).
[38] Basit, H., Ali, C. D. M. & Madhani, N. B. Erb Palsy. in StatPearls (StatPearls Publishing, 2022).
[39] Lipscomb, K. R., Gregory, K. & Shaw, K. The outcome of macrosomic infants weighing at least 4500 grams: Los Angeles County + University of Southern California experience. Obstet Gynecol 85, 558–564 (1995).
[40] Raio, L. et al. Perinatal outcome of fetuses with a birth weight greater than 4500 g: an analysis of 3356 cases. Eur J Obstet Gynecol Reprod Biol 109, 160–165 (2003).
[41] Neonatal Brachial Plexus Palsy | ACOG. https://www.acog.org/clinical/clinical-guidance/task-force-report/articles/2014/neonatal-brachial-plexus-palsy.
[42] King, J. R., Korst, L. M., Miller, D. A. & Ouzounian, J. G. Increased composite maternal and neonatal morbidity associated with ultrasonographically suspected fetal macrosomia. J Matern Fetal Neonatal Med 25, 1953–1959 (2012).
[43] Bryant, D. R., Leonardi, M. R., Landwehr, J. B. & Bottoms, S. F. Limited usefulness of fetal weight in predicting neonatal brachial plexus injury. Am J Obstet Gynecol 179, 686–689 (1998).
[44] Gillean, J. R., Coonrod, D. V., Russ, R. & Bay, R. C. Big infants in the neonatal intensive care unit. Am J Obstet Gynecol 192, 1948–1953; discussion 1953-1955 (2005).
[45] Cnattingius, S., Villamor, E., Lagerros, Y. T., Wikström, A.-K. & Granath, F. High birth weight and obesity--a vicious circle across generations. Int J Obes (Lond) 36, 1320–1324 (2012).
[46] Sparano, S. et al. Being macrosomic at birth is an independent predictor of overweight in children: results from the IDEFICS study. Matern Child Health J 17, 1373–1381 (2013).
[47] Shoulder Dystocia | ACOG. https://www.acog.org/en/clinical/clinical-guidance/practice-bulletin/articles/2017/05/shoulder-dystocia.
[48] Management of Stillbirth | ACOG. https://www.acog.org/en/clinical/clinical-guidance/obstetric-care-consensus/articles/2020/03/management-of-stillbirth.
[49] Labor Induction | ACOG. https://www.acog.org/en/womens-health/faqs/labor-induction.
[50] Combs, C. A., Singh, N. B. & Khoury, J. C. Elective induction versus spontaneous labor after sonographic diagnosis of fetal macrosomia. Obstet Gynecol 81, 492–496 (1993).

[51] Friesen, C. D., Miller, A. M. & Rayburn, W. F. Influence of spontaneous or induced labor on delivering the macrosomic fetus. Am J Perinatol 12, 63–66 (1995).
[52] Weeks, J. W., Pitman, T. & Spinnato, J. A. Fetal macrosomia: does antenatal prediction affect delivery route and birth outcome? Am J Obstet Gynecol 173, 1215–1219 (1995).
[53] Gonen, O. et al. Induction of labor versus expectant management in macrosomia: A randomized study. Obstetrics & Gynecology 89, 913–917 (1997).
[54] Leaphart, W. L., Meyer, M. C. & Capeless, E. L. Labor induction with a prenatal diagnosis of fetal macrosomia. J Matern Fetal Med 6, 99–102 (1997).
[55] Cheng, Y. W., Sparks, T. N., Laros, R. K., Nicholson, J. M. & Caughey, A. B. Impending macrosomia: will induction of labour modify the risk of caesarean delivery? BJOG 119, 402–409 (2012).
[56] Vendittelli, F., Rivière, O., Neveu, B. & Lémery, D. Does induction of labor for constitutionally large-for-gestational-age fetuses identified in utero reduce maternal morbidity? BMC Pregnancy and Childbirth 14, 156 (2014).
[57] Gross, S. J., Shime, J. & Farine, D. Shoulder dystocia: predictors and outcome. A five-year review. Am J Obstet Gynecol 156, 334–336 (1987).
[58] Langer, O., Berkus, M. D., Huff, R. W. & Samueloff, A. Shoulder dystocia: should the fetus weighing greater than or equal to 4000 grams be delivered by cesarean section? Am J Obstet Gynecol 165, 831–837 (1991).
[59] Delpapa, E. H. & Mueller-Heubach, E. Pregnancy outcome following ultrasound diagnosis of macrosomia. Obstet Gynecol 78, 340–343 (1991).
[60] Boulet, S. L., Alexander, G. R., Salihu, H. M. & Pass, M. Macrosomic births in the united states: determinants, outcomes, and proposed grades of risk. Am J Obstet Gynecol 188, 1372–1378 (2003).
[61] Overland, E. A., Vatten, L. J. & Eskild, A. Risk of shoulder dystocia: associations with parity and offspring birthweight. A population study of 1 914 544 deliveries. Acta Obstet Gynecol Scand 91, 483–488 (2012).
[62] Sheiner, E. et al. Determining factors associated with shoulder dystocia: a population-based study. Eur J Obstet Gynecol Reprod Biol 126, 11–15 (2006).
[63] Dall'Asta, A., Ghi, T., Pedrazzi, G. & Frusca, T. Does vacuum delivery carry a higher risk of shoulder dystocia? Review and meta-analysis of the literature. Eur J Obstet Gynecol Reprod Biol 204, 62–68 (2016).
[64] Operative Vaginal Birth | ACOG. https://www.acog.org/en/clinical/clinical-guidance/practice-bulletin/articles/2020/04/operative-vaginal-birth.
[65] Laughon, S. K. et al. Neonatal and maternal outcomes with prolonged second stage of labor. Obstet Gynecol 124, 57–67 (2014).
[66] Evans, A. & Refrow-Rutala, D. Medico-legal education: a pilot curriculum to fill the identified knowledge gap. J Grad Med Educ 2, 595–599 (2010).
[67] Wei, S. et al. Early amniotomy and early oxytocin for prevention of, or therapy for, delay in first stage spontaneous labour compared with routine care. Cochrane Database Syst Rev CD006794 (2013) doi:10.1002/14651858.CD006794.pub4.
[68] Sagi-Dain, L. & Sagi, S. The role of episiotomy in prevention and management of shoulder dystocia: a systematic review. Obstet Gynecol Surv 70, 354–362 (2015).
[69] Gherman, R. B., Tramont, J., Muffley, P. & Goodwin, T. M. Analysis of McRoberts' maneuver by x-ray pelvimetry. Obstet Gynecol 95, 43–47 (2000).
[70] Gonik, B., Stringer, C. A. & Held, B. An alternate maneuver for management of shoulder dystocia. Am J Obstet Gynecol 145, 882–884 (1983).
[71] Cluver, C. A. & Hofmeyr, G. J. Posterior axilla sling traction for shoulder dystocia: case review and a new method of shoulder rotation with the sling. American Journal of Obstetrics & Gynecology 212, 784.e1-784.e7 (2015).
[72] Bruner, J. P., Drummond, S. B., Meenan, A. L. & Gaskin, I. M. All-fours maneuver for reducing shoulder dystocia during labor. J Reprod Med 43, 439–443 (1998).
[73] Goffman, D., Heo, H., Pardanani, S., Merkatz, I. R. & Bernstein, P. S. Improving shoulder dystocia management among resident and attending physicians using simulations. Am J Obstet Gynecol 199, 294.e1–5 (2008).
[74] Goffman, D., Heo, H., Chazotte, C., Merkatz, I. R. & Bernstein, P. S. Using simulation training to improve shoulder dystocia documentation. Obstet Gynecol 112, 1284–1287 (2008).
[75] Deering, S., Poggi, S., Macedonia, C., Gherman, R. & Satin, A. J. Improving resident competency in the management of shoulder dystocia with simulation training. Obstet Gynecol 103, 1224–1228 (2004).

[76] Crofts, J. F. et al. Management of shoulder dystocia: skill retention 6 and 12 months after training. Obstet Gynecol 110, 1069–1074 (2007).
[77] Crofts, J. F. et al. Observations from 450 shoulder dystocia simulations: lessons for skills training. Obstet Gynecol 112, 906–912 (2008).
[78] Hunt, E. A., Shilkofski, N. A., Stavroudis, T. A. & Nelson, K. L. Simulation: translation to improved team performance. Anesthesiol Clin 25, 301–319 (2007).
[79] Macrosomia | ACOG. https://www.acog.org/clinical/clinical-guidance/practice-bulletin/articles/2020/01/macrosomia.
[80] 1.Wiebe, H. W., Boulé, N. G., Chari, R. & Davenport, M. H. The effect of supervised prenatal exercise on fetal growth: a meta-analysis. Obstet Gynecol 125, 1185–1194 (2015).
[81] Seijmonsbergen-Schermers, A. E. et al. Variations in use of childbirth interventions in 13 high-income countries: A multinational cross-sectional study. PLOS Medicine 17, e1003103 (2020).
[82] Ganchimeg, T. et al. Pregnancy and childbirth outcomes among adolescent mothers: a World Health Organization multicountry study. BJOG: An International Journal of Obstetrics & Gynaecology 121, 40–48 (2014).
[83] Nagata, C. et al. Complications and adverse outcomes in pregnancy and childbirth among women who conceived by assisted reproductive technologies: a nationwide birth cohort study of Japan environment and children's study. BMC Pregnancy Childbirth 19, 77 (2019).
[84] Bowers, K. et al. Gestational diabetes, pre-pregnancy obesity and pregnancy weight gain in relation to excess fetal growth: variations by race/ethnicity. Diabetologia 56, 1263–1271 (2013).
[85] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine 23, 89–109 (2001).
[86] Afzal, N. et al. Natural language processing of clinical notes for identification of critical limb ischemia. Int J Med Inform 111, 83–89 (2018).
[87] Gulshan, V. et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 316, 2402–2410 (2016).
[88] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).
[89] Hinton, G. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine 29, 82–97 (2012).
[90] Shamout, F., Zhu, T. & Clifton, D. A. Machine Learning for Clinical Outcome Prediction. IEEE Reviews in Biomedical Engineering 14, 116–126 (2021).
[91] von Dadelszen, P. et al. Prediction of adverse maternal outcomes in pre-eclampsia: development and validation of the fullPIERS model. Lancet 377, 219–227 (2011).
[92] de Arriba-Pérez, F., Caeiro-Rodríguez, M. & Santos-Gago, J. M. Collection and Processing of Data from Wrist Wearable Devices in Heterogeneous and Multiple-User Scenarios. Sensors (Basel) 16, 1538 (2016).
[93] Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data 6, 54 (2019).
[94] Praveena, M. & Jaiganesh, V. A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. International Journal of Computer Applications 169, 32–35 (2017).
[95] Ghahramani, Z. Unsupervised Learning. in Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures (eds. Bousquet, O., von Luxburg, U. & Rätsch, G.) 72–112 (Springer, 2004). doi:10.1007/978-3-540-28650-9_5.
[96] Li, Y. Deep Reinforcement Learning: An Overview. arXiv:1701.07274 [cs] (2018).
[97] Chen, L. Overview of clinical prediction models. Ann Transl Med 8, 71 (2020).
[98] Lee, Y., Bang, H. & Kim, D. J. How to Establish Clinical Prediction Models. Endocrinol Metab (Seoul) 31, 38–44 (2016).
[99] Cole, T. J. Applied logistic regression. D. W. Hosmer and S. Lemeshow, Wiley, New York, 1989. No. of pages: xiii + 307. Price: £36.00. Statistics in Medicine 10, 1162–1163 (1991).
[100] Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology 110, 12–22 (2019).
[101] Cao, H. et al. Predicting ICU hemodynamic instability using continuous multiparameter trends. Annu Int Conf IEEE Eng Med Biol Soc 2008, 3803–3806 (2008).

[102] Loekito, E. et al. Common laboratory tests predict imminent death in ward patients. Resuscitation 84, 280–285 (2013).
[103] Churpek, M. M., Yuen, T. C. & Edelson, D. P. Predicting clinical deterioration in the hospital: the impact of outcome selection. Resuscitation 84, 564–568 (2013).
[104] Bewick, V., Cheek, L. & Ball, J. Statistics review 14: Logistic regression. Critical Care 9, 112 (2005).
[105] Emmert-Streib, F. & Dehmer, M. High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. MAKE 1, 359–383 (2019).
[106] Klau, S., Jurinovic, V., Hornung, R., Herold, T. & Boulesteix, A.-L. Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. BMC Bioinformatics 19, 322 (2018).
[107] Tsur, A. et al. Development and validation of a machine-learning model for prediction of shoulder dystocia. Ultrasound Obstet Gynecol 56, 588–596 (2020).
[108] Schmidhuber, J. Deep learning in neural networks: An overview. Neural Networks 61, 85–117 (2015).
[109] Hu, S. B., Wong, D. J. L., Correa, A., Li, N. & Deng, J. C. Prediction of Clinical Deterioration in Hospitalized Adult Patients with Hematologic Malignancies Using a Neural Network Model. PLoS One 11, e0161401 (2016).
[110] Heo, J. et al. Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. Stroke 50, 1263–1265 (2019).
[111] Geurts, P., Irrthum, A. & Wehenkel, L. Supervised learning with decision tree-based methods in computational and systems biology. Mol Biosyst 5, 1593–1605 (2009).
[112] Hanko, M. et al. Random Forest-Based Prediction of Outcome and Mortality in Patients with Traumatic Brain Injury Undergoing Primary Decompressive Craniectomy. World Neurosurg 148, e450–e458 (2021).
[113] Deist, T. M. et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys 45, 3449–3459 (2018).
[114] Fernandez-Lozano, C. et al. Random forest-based prediction of stroke outcome. Sci Rep 11, 10071 (2021).
[115] Choi, E., Bahadori, M. T., Searles, E., Coffey, C. & Sun, J. Multi-layer Representation Learning for Medical Concepts. arXiv:1602.05568 [cs] (2016).
[116] Gherman, R. B. et al. Shoulder dystocia: the unpreventable obstetric emergency with empiric management guidelines. Am J Obstet Gynecol 195, 657–672 (2006).
[117] Zhang, M.-L., Li, Y.-K., Yang, H. & Liu, X.-Y. Towards Class-Imbalance Aware Multi-Label Learning. IEEE Trans Cybern PP, (2020).
[118] Mirkes, E. M., Coats, T. J., Levesley, J. & Gorban, A. N. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. Comput Biol Med 75, 203–216 (2016).
[119] Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 20, 117–121 (2013).
[120] Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20, 144–151 (2013).
[121] Forbes, A. D. Classification-algorithm evaluation: five performance measures based on confusion matrices. J Clin Monit 11, 189–206 (1995).
[122] Lever, J., Krzywinski, M. & Altman, N. Classification evaluation. Nature Methods 13, 603–604 (2016).
[123] Taylor, J. R. & Taylor, S. L. in L. J. R. Introduction To Error Analysis: The Study of Uncertainties in Physical Measurements. (University Science Books, 1997).
[124] Encyclopedia of Machine Learning. (Springer US, 2010). doi:10.1007/978-0-387-30164-8.
[125] Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006).
[126] Powers, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach. Learn. Technol. 2, (2008).
[127] Rubinsteyn, A. iskandr/fancyimpute. (2022). https://github.com/iskandr/fancyimpute
[128] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
[129] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
[130] Martins, D. XGBoost: A Complete Guide to Fine-Tune and Optimize your Model. Medium https://towardsdatascience.com/xgboost-fine-tune-and-optimize-your-model-23d996fab663 (2021).
[131] XGBoost Parameters — xgboost 1.6.1 documentation. https://xgboost.readthedocs.io/en/stable/parameter.html.

[132] fernando. Bayesian Optimization. (2022). https://github.com/fmfn/BayesianOptimization.
[133] Graphviz. Graphviz https://graphviz.org/.
[134] SHAP Values. https://kaggle.com/dansbecker/shap-values.
[135] Bragança, H., Colonna, J. G., Oliveira, H. A. B. F. & Souto, E. How Validation Methodology Influences Human Activity Recognition Mobile Systems. Sensors (Basel) 22, 2360 (2022).
[136] Budholiya, K., Shrivastava, S. K. & Sharma, V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University - Computer and Information Sciences (2020) doi:10.1016/j.jksuci.2020.10.013.
[137] Donahoe, J. T. & McGuire, T. G. The vexing relationship between socioeconomic status and health. Israel Journal of Health Policy Research 9, 68 (2020).
[138] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. http://arxiv.org/abs/1705.09406 (2017) doi:10.48550/arXiv.1705.09406.

Supplementary Material

Imputation	Class imbalance	Feature Selection	Optimization	Experiment	AUROC	AUPRC	F1 score
fancyimpute	No	No	Grid search	1	0.792	0.786	0.708
			Bayesian	2	0.8	**0.817**	0.695
		Yes	Grid search	3	0.811	0.813	0.707
			Bayesian	4	0.786	0.792	0.68
	Yes	No	Grid search	5	0.735	0.026	0.039
			Bayesian	6	0.728	0.029	0
		Yes	Grid search	7	0.703	0.097	0.031
			Bayesian	8	0.722	0.039	0
XGB missing param.	No	No	Grid search	9	0.811	**0.817**	0.651
			Bayesian	10	0.77	0.758	0.72
		Yes	Grid search	11	**0.812**	0.804	**0.734**
			Bayesian	12	0.781	0.767	0.673
	Yes	No	Grid search	13	0.749	0.031	0.046
			Bayesian	14	0.736	0.033	0.037
		Yes	Grid search	15	0.703	0.135	0.029
			Bayesian	16	0.745	0.064	0

Table 2. Summary of results from all experiments. Highest values in each evaluation metric column are highlighted.

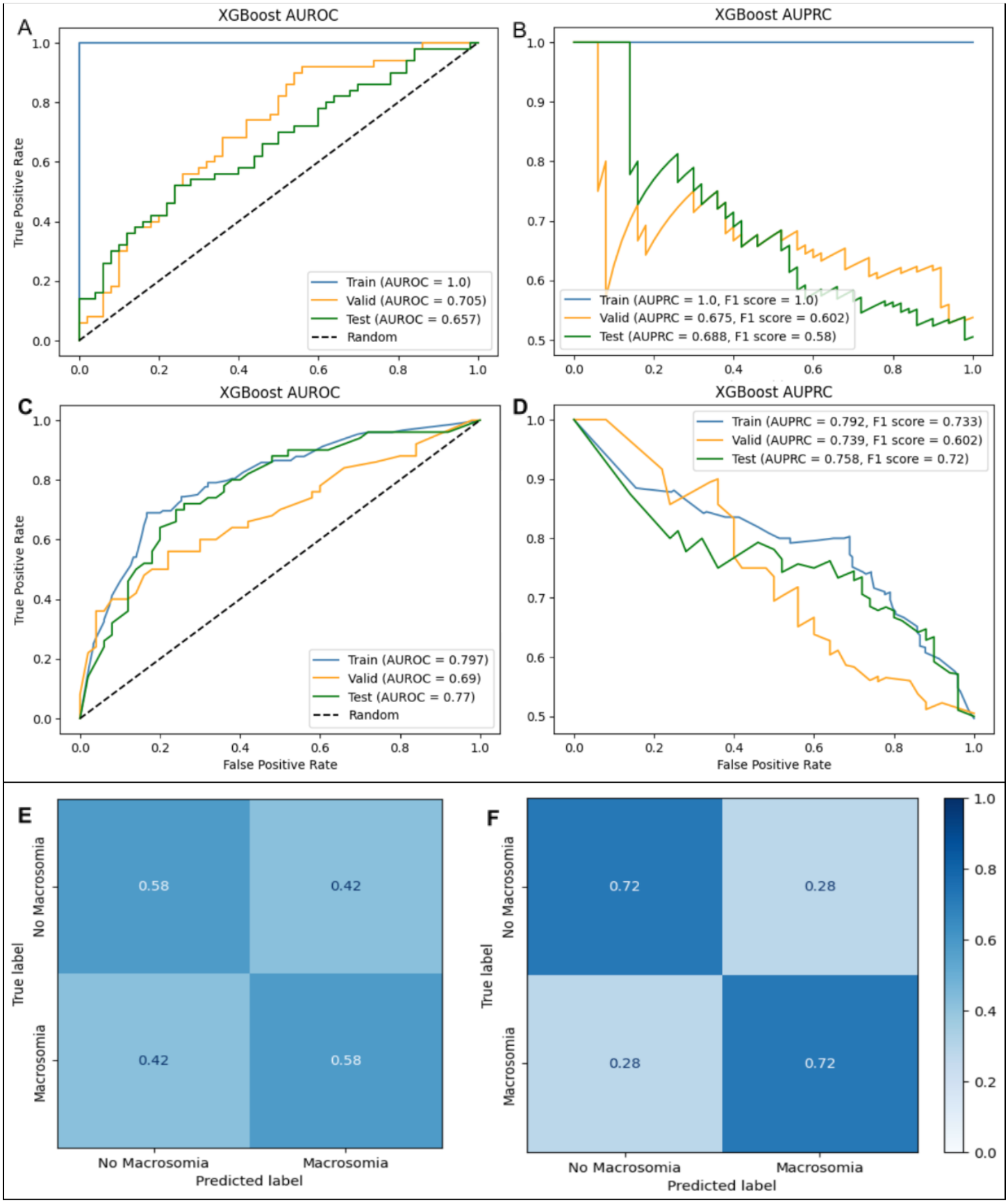

Figure 7. Results from experiment 10. AUROC, AUPRC, and confusion matrices were plotted before and after optimization for all experiments; the experiment with the second highest F1 score is shown here. **A**, **B**, and **E** show plots before optimization; **C**, **D**, and **F** show plots after optimization of hyperparameters.