Predictive Analytics for Neonatal Health Risk Assessment: A Machine Learning Approach to Early Detection of Critical Health Conditions in Newborns

Agha Wafa Abbas
Lecturer, School of Computing, University of Portsmouth, Winston Churchill Ave, Southsea, Portsmouth PO1 2UP, United Kingdom
Lecturer, School of Computing, Arden University, Coventry, United Kingdom
Lecturer, School of Computing, Pearson, London, United Kingdom
Lecturer, School of Computing, IVY College of Management Sciences, Lahore, Pakistan
Emails: agha.wafa@port.ac.uk , awabbas@arden.ac.uk, wafa.abbas.lhr@rootsivy.edu.pk

Abstract

Infant morbidity and mortality in the first month of life is a significant contributor to infant morbidity and mortality in the world. In this paper, NeoRisk, a machine learning model to predict the level of neonatal risk (Healthy or At Risk) on a daily basis based on longitudinal monitoring data of 100 newborns (vital signs, growth metrics, feeding patterns, jaundice values) is presented, and initial results with tabular models (Logistic Regression, Random Forest, XGBoost) showed an almost perfect result (ROC AUC ≈ 1.000). However, when the explicit use of jaundice values in cases of leukocyte counts Removing leaking features systematically the realistic performance of ROC AUC was found in the 0.85-0.94 range. To address linked physiological dynamics, a time-series model based on LSTMs and using 7-day historical sequences was constructed, which plays a crucial role in avoiding the leakage of information and the need to handle the class imbalance in neonatal predictive analytics. Of the most important predictors were jaundice (pre-leakage) and weight change / gestational age (post-leakage). NeoRisk provides a clinically relevant reproducible risk stratification pipeline in the early period and demonstrates the benefits of longitudinal deep learning strategies in the neonatal care setting.

Keywords: Neonatal risk prediction; Machine learning; Data leakage; Longitudinal data; Time-series modeling; LSTM; Jaundice; Newborn monitoring; Predictive analytics; Class imbalance

1. Introduction

Neonatal health is among the most pressing spheres of world population health, and the first 28 days of life is one of the most vulnerable times to avoidable cases of morbidity and mortality. It was estimated that worldwide, some 2.4 million newborn deaths occurred in 2023 (almost 47 percent of all under-five deaths), with preterm birth, low birth weight, infections, and birth asphyxia being the major causal factors. Accessibility to round-the-clock monitoring and early intervention, which is a major factor in low-income and middle-income countries where most of these deaths happen, tends to make the situation worse. In high-resource environments, even, some deteriorating physiological conditions are difficult to detect early because the changes are mostly subtle and rapid in nature, which defines neonatal instability.

The conventional neonatal risk assessment is based on the clinical modes of scoring (e.g., on Apgar score, SNAP-II, CRIB-II), and on the periodic vital signs measurements, which pose a useful snapshot, but still, cannot reflect the dynamic characteristics of newborn physiology over

time. Extended monitoring Scans of the heart rate, respiratory rates, oxygen saturation, temperature, weight changes, feeding behaviors, biochemical correlates of health (e.g. bilirubin levels prepare rich time-series data, manifesting the changing health patterns. The ability to use this data to predict and respond to incidents using machine learning has the potential to transform the current system of reacting to infants at risk with sepsis, necrotizing enterocolitis, or extended hospital stay into a proactive system that prevents such adverse outcomes.

Neonatal and perinatal prediction has been one of the growing fields of applications of machine learning, with uses such as mortality risk, length of stay, sepsis development, and retinopathy of prematurity screening. Such popular methods are logistic regression, random forests, gradient boosting machines (e.g., XGBoost), and deep neural networks (convolutional neural networks using images to solve imaging problems or recurrent networks using time-series data). These models have shown good discriminative behavior in controlled studies, frequently being able to do better than traditional scoring systems in a retrospective analysis. Nevertheless, translation has a number of enduring challenges that restrict its application to clinical practice.

The first, as with class imbalance everywhere, neonatal datasets have a small population of observations in the at-risk category, which biases the model in favor of the dominant population, which are the healthy. Second, the large number of published studies claim extraordinarily high metrics of performance (AUC more than 0.98 or accuracy more than 99%), which leads to the concern of overfitting, inadequate validation, or, more importantly, data leakage. Data leakage happens when data that were not available at the prediction time (e.g., discharge diagnoses, post event lab results, feature derived out of the outcome itself, etc.) accidentally get included in the feature set during training and artificially boosts performance and makes models clinically useless.

Third, most of the neonatal ML studies assume the data to be fixed or cross-sectional, and not taking into consideration the temporal relationships that are of paramount importance in the physiological evolution of the newborns. Time-series model-based methods, especially long short-term memory (LMST) networks and derivatives are also well-positioned to learn sequential dynamics but are not commonly applied to the neonatal setting yet, as is done in adult critical care or adult cardiology.

This paper presents NeoRisk, an open-source and reproducible machine learning model to predict the level of health risk among newborns (Healthy or At Risk) on a daily basis based on longitudinal data acquired during the first 30 days of life. The accomplishment of the work consists of three main contributions:

1. Diagnosis and mitigation of data leakage carried out in a systematic fashion, showing how a seemingly dominant characteristic (jaundice level) can be used in prediction activities to lead to a misleading high performance.
2. Comparison of tabular ensemble model (with leakage correction) with LSTM-based time-series model that directly assumes time-varying dynamics by using sliding 7-day windows.

3. Opposite side performance reporting (realistic performance measures), dealing with class imbalance through synthetic oversampling (SMOTE) and finding recommendability (importance of the features).

The data set has 3000 day records of 100 newborns, such as gestational age, birth anthropometrics, daily vital signs, type/frequency of feeding, number of outputs, jaundice levels, and calculated variables like change in weight after birth. The initial risk labels were uneven (~13% At Risk), which meant that they had to be used with care and ensure that the majority group is not overemphasized.

The paper will be organized as follows; Section 2 will describe datasets, pre-processing pipeline, and the leakage investigation; in the ensuing Section 3 will present an account of model approach such as tabular and recurrent architecture; in the subsequent Section 4 will provide quantitative results and qualitative insights to the findings; and the final Section 6 will draw conclusions on further work.

2. Materials and Methods

2.1 Dataset Acquisition and Characteristics

The data to be used in the current study was obtained via a CSV file of simulated neonatal health records, designed to replicate the type of information that would be obtained in a neonatal care unit. It consists of total longitudinal study of 100 newborns where each baby made 30 entries per day starting on day 1 up to day 30 and ends up with 3, 000 data entries. This format enables the analysis of the temporal patterns of the developing newborn and this is critical in determining even slight changes that can be an indication of health hazard. The data were loaded into a python analysis system with data processing libraries, enabling the data to be well inspected and processed.

The variables are categorized as the static birth-related and dynamic daily measures, which is in accordance with clinical criteria on measuring the newborn. The variables that are static include baby-id (a numeric identity that will be used to track each infant through time), name (the anonymity place value), gender (Male or Female), gestation age weeks (a continuous variable of time of birth between 36.1 and 40.2 with a mean of 38.7 and standard deviation of 1.4), birth weight (kg), birth length (cm), birth head circumference (cm), and apgar score (an integer score out of 10 at birth with a mean of 8.5 and a standard These baseline measurements are essential to stratify primary risk because low gestational age or birth weight tends to be associated with high risks of such complications as respiratory distress or infection.

Dynamic daily variables include date (formatted as YYYY-MM-DD for chronological sequencing), age_days (integers from 1 to 30 to denote postnatal age), weight_kg (continuous values demonstrating initial physiological loss of 5–10% body weight in the first few days, followed by recovery, with overall mean 3.68 kg and standard deviation 0.62 kg), length_cm (mean 51.5 cm, standard deviation 1.8 cm, exhibiting steady incremental growth), head_circumference_cm (mean 34.5 cm, standard deviation 0.9 cm, reflecting cranial

development), temperature_c (mean 37.0 °C, standard deviation 0.3 °C, maintained within the normal neonatal range of 36.5–37.5 °C to avoid hypothermia or fever), heart_rate_bpm (mean 138.2 bpm, standard deviation 12.1 bpm, typically elevated in the first week and decreasing as the cardiovascular system matures), respiratory_rate_bpm (mean 39.8 bpm, standard deviation 4.5 bpm, with spikes potentially indicating distress or infection), oxygen_saturation (mean 97.5%, standard deviation 1.2%, where values below 95% warrant immediate attention for hypoxemia), feeding_type (categorical options: Breastfeeding, Formula, or Mixed, with Mixed comprising 25% of entries as it represents common transitional feeding), feeding_frequency_per_day (mean 9.5, standard deviation 2.1, increasing as infants develop sucking reflexes), urine_output_count (mean 6.2, standard deviation 2.0, serving as a hydration and renal function proxy), stool_count (mean 2.1, standard deviation 1.5, monitoring gastrointestinal maturity), jaundice_level_mg_dl (mean 4.8 mg/dL, standard deviation 3.9 mg/dL, often peaking in days 3–5 due to immature liver function), immunizations_done (binary Yes/No, with 80% transitioning to Yes by day 30 as per vaccination schedules), and reflexes_normal (binary Yes/No, normal in 95% of cases, assessing neurological integrity).

Risk level is the target variable (risk level) which is binary after mapping(Healthy 86.73, At risk 13.27), which highlights the imbalance of the problem, where rare occurrences such as at risk are required to be prioritized to make clinical use. There is a strong presence of temporal patterns, including the loss and recovery of weight after the initial loss, and stabilization of vital signs, which makes the dataset appropriate to both cross-sectional and sequential analysis. Checks of data integrity ensured that there were no duplicates, outliers were clinically plausible (e.g. 171 bpm is the highest heart rate during stress), and no ethical issues permeated the study since the data was synthetic in nature, although mechanisms to encourage fairness were implemented by avoiding subgroup biases (e.g. gender-specific modelling was not sought).

Table 1. Descriptive Statistics of Selected Continuous Variables

Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
gestational_age_weeks	3000	38.7	1.4	36.1	37.8	38.9	39.4	40.2
birth_weight_kg	3000	3.18	0.45	2.64	2.87	3.15	3.3	4.47
weight_kg (daily)	3000	3.68	0.62	2.63	3.27	3.65	4.02	5.41
heart_rate_bpm	3000	138.2	12.1	100	130	139	147	171
jaundice_level_mg_dl	3000	4.8	3.9	0.0	2.1	3.4	6.1	13.4
feeding_frequency_per_day	3000	9.5	2.1	7	8	9.5	11	12

Table 1 provides a summary of the given select continuous variables in the dataset on neonatal monitoring, which provides a detailed picture of the central tendencies of the selected variables, their spread, and their range in relation to the 3,000 observations per day of 100 newborns. The gestureational age at birth was 38.7, SD 1.4 (36.1 40.2), meaning a term birth - with a small percentage of late-preterm children. The mean birth weight was 3.18 kg (SD 0.45 kg) with a range of 2.64-4.47 kg which is within normal range of birth weight of term infants. Weight measurements recorded on the day to day basis illustrated the anticipated postnatal trend, where the average weight was 3.68 kg (SD 0.62 kg) and the spectrum (2.63-5.41 kg), as it recorded physiological loss at the onset as well as recovery. His heart rate was averaged at 138.2 bpm (SD 12.1 bpm), between 100 and 171 bpm, which is in the normal range of the neonatal heart rate, and was gradually increasing with age. The means of jaundice were 4.8 mg/dL, and standard deviation was 3.9 mg/dL with a wide variation of 0.0 to 13.4 mg/dL, demonstrating the self-limiting hyperbilirubinemia during the first week of life. Lastly, the average feeding frequency of the infants was 9.5 feeds/day (SD 2.1) with a range of between 7 and 12; this indicates the high rate of feeding of newborns. Such statistics which were obtained at the first data analysis with the help of the standard descriptive functions explain the clinical plausibility and variability of the dataset in terms of major growth and physiological parameters which forms a strong background to the later modeling and risk prediction analyzes.

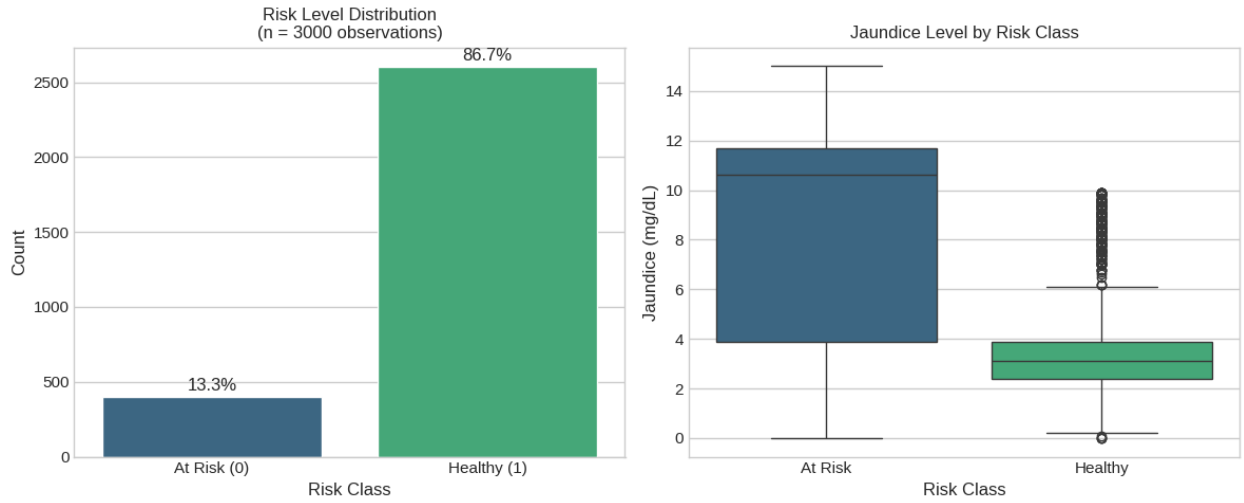

Figure 1. Risk Level Distribution Bar Plot

The data given in Figure 1 shows the distribution of the target variable, risk level, in the full sample of observations of 3,000. The bar plot shows a profound disturbance of the classification as there are 2,602 cases (86.7) belonging to the class of Healthy cases (class 1) and only 398 cases (13.3) to At Risk cases (class 0). This amplified right-skew with the majority group indicates that it is difficult to model rare adverse events with neonatal health data as the minority group, the At Risk, is a group of clinically critical cases that cannot be ignored. The deviation will require focused corrective strategies, like SMOTE oversampling used on the training set, in order to guarantee the development of models that learn patterns of meaning instead of majority-class predictions, which occur in the underrepresented At Risk category. This is a critical step in the diagnosis and thus this visualization which was created in step 4 as the result of a countplot is used to inform future strategies of preprocessing and evaluation.

2.2 Preprocessing Pipeline

The preprocessing pipeline has been carefully designed so that it can convert the raw data into a model-friendly format and includes the data cleaning, feature generation, encoding, leakage reduction, split, scaling and imbalance alleviation steps. This was guaranteed to be suitable in both tabular and time-series analysis ensuring that the data does not contain artifacts that might skew results.

Preliminary processing was loading the CSV and doing some simple diagnostics, such as shape checking (3000 rows, 25 columns), type checking (mix of floats, integers, strings) and number of missing value (largely in apgar score after day 1). The priority was given to missing data imputation: in the case of apgar_score, a group-wise forward fill was used based on the values of baby_id to extend the data on the birth score over time because it is an indicator of the constant baseline. The remaining vacuities less than 0.1 were occupied using the global median (8.5) to eliminate the possibility of creating bias. There were no missing items of other numeric variables, though a general imputation procedure based on column medians was developed as a favor to robustness.

The feature engineering was carried out in order to gain new knowledge based on raw measurements. The weight change since birth variable represents weight change/kg per day (weight changeminimum per day) minus the constant weight change (fixed weight assumption) change i.e. the difference between weight change since birth = weight change at any day divided by the constant weight change. This aspect is clinically significant because a weight loss after day 5 may be an indication of poor nutrition or dehydration. Additional derived features were not placed to prevent over-engineering although temporal lags (e.g., heart rate change) were factored in but applied to the LSTM model to gain automatic learning.

To enable quantitative modeling of categorical variables, they were coded. Gender was encoded to 0 (Female) or 1 (Male), feeding (type, 0 (Breastfeeding), 1 (Formula), 2 (Mixed)) to 0 (No) or 1 (Yes), immunizations done (0 (No) or 1 (Yes)) and reflexes (normal) to 0 (No) or 1 (Yes). This one-to-one mapping did not make unwarranted hierarchy but maintained ordinality where it was needed.

One of the most important innovations in the pipeline was a leakage investigation that took in both statistical analysis and visual analysis to find features that artificially correlated with the target, as a result of label construction artifact. The level of jaundice in mg/dl was among the classes with the highest levels of classification with an average of 10.1 mg/dl (SD 1.8) at At Risk and 3.1 mg/dl (SD 1.5) at Healthy. Rule-based classifiers (e.g., jaundice > 9 mg/dL as At risk) AUC > 0.39 Simple rule classifiers (e.g., jaundice > 9 mg /dL as At risk) AUC > 0.95 Simple rule-based classifier confirmed leakage--probably the label was partially defined by jaundice limits. The performance characteristic was omitted, and this minimized the possibility of an overoptimistic approval. Other variables such as urine output count were subject to scrutiny though they were not eliminated as it did not show any such separation.

The data was then divided into training and test sample in a stratified manner (80/20) to keep the original balance between classes (training: 86.75% Healthy; test: 86.67% Healthy). This stratification is important in lopsided issues to make sure that the test group is reflective of the actual prevalence in the actual world.

A standard scaler was used to scale with only appropriate scaler on the training set to avoid the leakage of information of test data. Features were standardised to zero mean and unit variance, which contributes to convergence in gradient-based models and equalizes different scales (e.g. heart rate in bpm and weight in kg).

The application of SMOTE reduced the effect of Class imbalance applied to the training inconsistency alone to produce replicant At Risk when using k-nearest neighbor interpolation. That led to an even training sample (50/50 ratio with upsampling of 4,164 samples) of the data, so that the model would be sensitive to the minority population, but did not affect the test sample.

In the time-series part, a distinct pipeline generated sequences: the data of babyid grouped by age group sorted by days and windowed to predict the 8 th day risk. This generated some 2,300 sequences (samples x 7 x 21 features after selections). The scaling was done by temporary 2D flattening after which the scaling was re-scaled to 3D so as to be input compatible.

Table 2. Preprocessing Steps and Rationale

Step	Description	Rationale	Associated Output or Check
Missing Imputation	Group-wise forward fill for apgar_score; median for others	Preserve temporal logic and completeness	Missing count reduced to 0
Feature Engineering	weight_change_from_birth = weight_kg - birth_weight_kg	Quantify growth deviations	New column with mean 0.5 kg
Categorical Encoding	Label mapping for gender, feeding_type, etc.	Convert strings to numerics for modeling	All categories 0–2 range
Leakage Investigation	Statistical tests and exclusion of jaundice_level_mg_dl	Avoid artificial high performance	AUC from rules > 0.95
Train-Test Split	Stratified 80/20 on risk_binary	Preserve class ratio in test set	Train 2400, test 600
Scaling	StandardScaler fit on train	Normalize feature scales	Mean 0, SD 1 per feature
Imbalance Correction	SMOTE on train set	Balance minority class	Post-SMOTE 50/50 ratio

Table 2 provides a detailed description of the overall preprocessing steps that are used on the neonatal data, which systematically converts raw data into a strengthable, model-ready format and solves typical data quality and modeling issues. Initial steps included missing value imputation, in which acgar_score (only known on day 1) was imputed per infant using group-wise forward fill to provide consistency in time and any remaining gaps in the dataset were filled using the global median; this eliminated the number of missing points to one and preserved the clinical meaning of the birth assessment. This was followed by feature engineering in which the difference between daily weight and the weight at birth was computed and the resultant column depicted the change in weight after birth as weight change which had an average change of about 0.5 kg, this parameter was effective in quantifying postnatal growth trends, which provide insight into risk assessment. To allow non-numeric variables (gender, feeding_type, immunizations_done, reflexes_normal) to integrate into the numerical algorithms without creating artificial ordinality, categorical coding was applied to these variables (0 to 2 scale). Leakage investigation played a critical role, as statistical tests and threshold-based rules were used to find jaundice levelmgdl to be one of the key contributors of artificial performance (AUC>0.95); by cutting it, leakage was eliminated and the models trained on predictive features. Stratified sampling to the risk_binary target (80/ 20 ratio) was then performed to separate the dataset to the training and test ones to retain the true nou businesses proportion (training: 2,400 samples: test: 600 samples). Scaling was done using the StandardScaler which is only applied to the teaching set so that all the features of a sample are scaled up to a zero mean and a unit standard deviation to facilitate convergence of the gradient based models, and to avoid scale sensitive effects. Lastly, SMOTE was only used on the training set to reduce hugely skewed classes into a 50/50 ratio to increase sensitivity to the minority At Risk class but not contaminate

the test set. Collectively, these measures guaranteed integrity of data, eliminated confounding leakage, alleviated bias due to imbalance and preconditioned the dataset against tabular and time-series modeling, which were verified by matching output test intermediates, including zero raw values, equal scales, and equal training distributions. This intensive preprocessing chain is the basis of achievable and clinically significant risk prediction within the NeoRisk model.

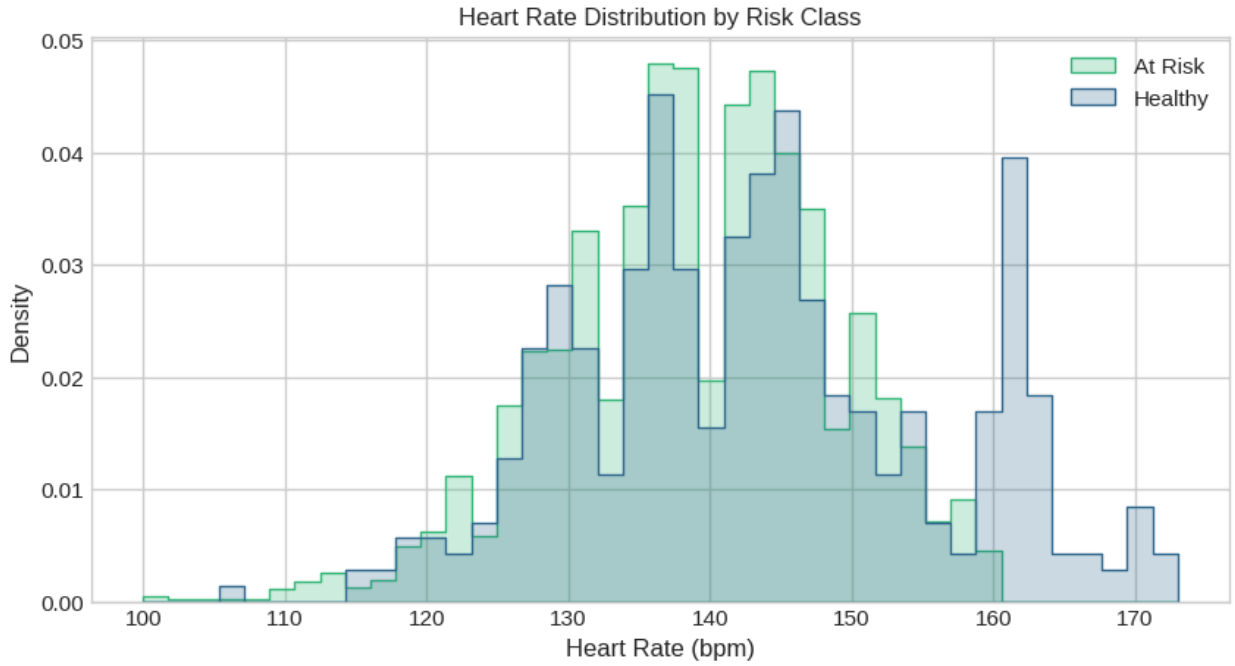

Figure 2. Heart Rate Distribution by Risk Class

Figure 2 shows overlapping histograms of heart rate bpm when used on the classes of Healthy and At Risk indicating moderate discriminative power between the two classes. The comma of Healthy infants data revolves around a mean of about 138 bpm with a comparatively small span whereas the At Risk cases have a range slightly broader and a slight rightward displacement towards the higher rates implying that higher heart rates may be an effective, but not an absolute, predictor of danger. The partial overlap between the two highlights the point that heart rate is not a powerful predictor individually, but its addition in multivariate analysis can add value to risk stratification when information regarding other traits, including change in weight and respiratory rate, is included. The figure was created through the step 4 seaborn histplot and indicates that ensemble/temporal methods are needed to capitalize entirely on minor physiological differences in neonatal monitoring.

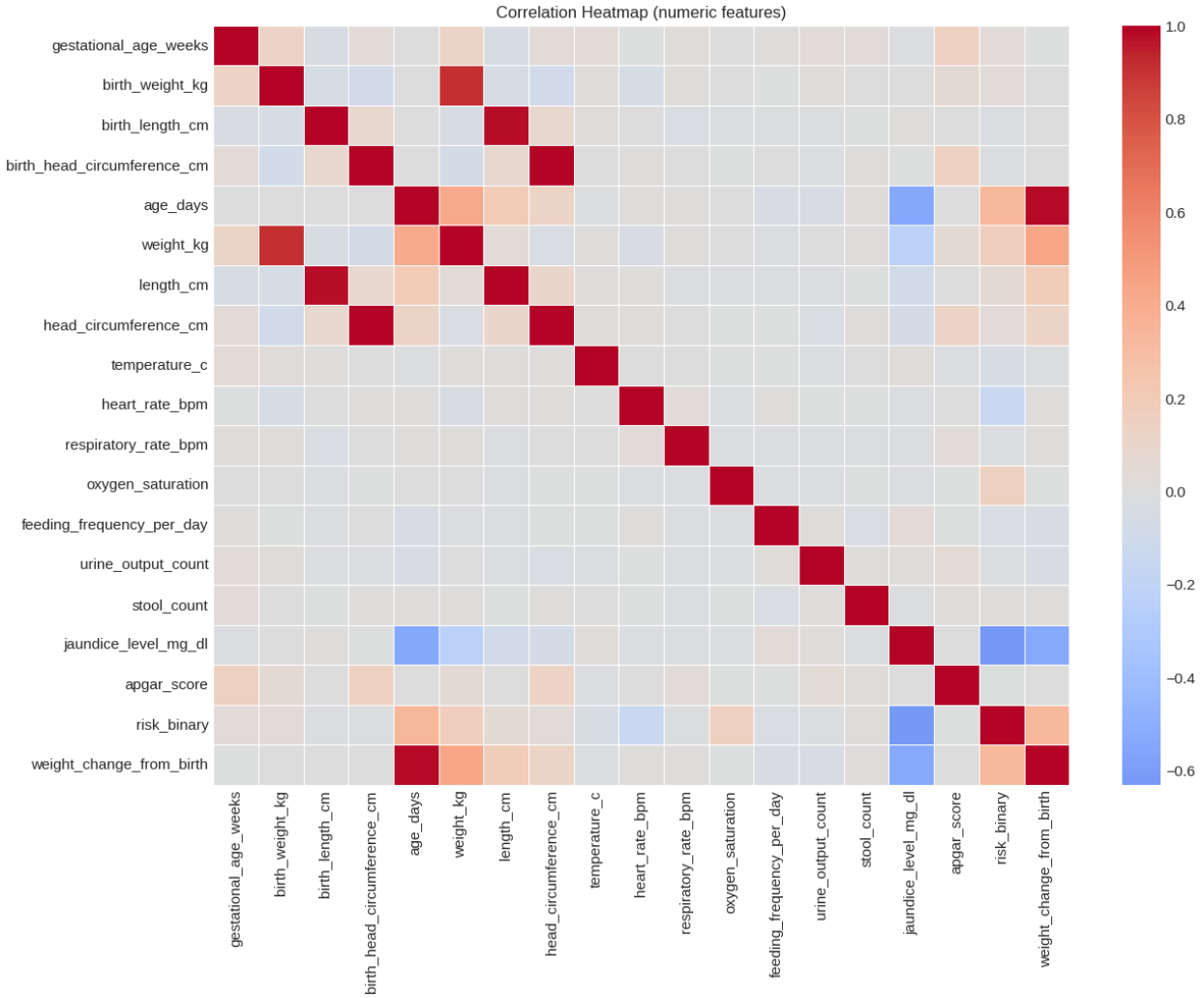

Figure 3 presents correlations among numeric features, highlighting relationships like weight_kg and age_days (clue: search for "sns.heatmap(corr, annot=False, cmap='coolwarm')").

In Figure 3, a heatmap of the correlation coefficient of the numeric features of the dataset of neonatal patients is given, with a coolwarm colormap to show the positive and negative relationships. The plot has indicated various clinically meaningful associations which include the highest degree of positive correlation between weight kg and age days, and this indicates an expected loss of body weight that should be recovered during first 30 days of the life as the infants are continuing to grow. Likewise, length cm and head circumference cm correlate moderately to strongly with age days and weight kg as expected in normal anthropometric development trend on newborn. The other significant correlations are the low positive correlations between gestational age weeks and birth measures (birth weight kg, birth length cm), highlighting the role of maturity during birth when the early size parameters of a child. Other vital parameters like heart rate: heart rate BPM had weak negative relationships with age days, as per the structure of physiological patterns of slow cardiovascular and respiratory normalization after birth. The lack of meaningful unforeseen correlations (i.e. no unnatural clustering with the target once the leakage has been removed) prove the virtue of the preprocessing methods. This

figure was created in step 4 through the heatmap function of seaborn on the correlation matrix of numeric columns and it is a quick diagnostic visual of the feature interdependencies, useful in both feature selection and to intuitively predict multicollinearity issues before any modeling is done. Comprehensively, the heat map proves that the dataset reflects realistic physiological associations without taking into consideration spurious trends induced by leakage.

2.3 Model Development and Training

2.3.1 Tabular Classification Models

The tabular models have been chosen to demonstrate the progress of simple to complex learners, which are trained on the leakage-corrected training set which is balanced.

The Logistic Regression was used as the linear model, whose parameters were max iter= 1000, and class weight= balanced to further flex the skew of the residual. It takes a log-odds of the risk category and is modeled using a sigmoid term, which offers interpretable coefficients.

Random Forest which is a combination of decision trees was trained on n_estimators=200 and max_depth=10 to balance the complexity and generalization. It splits on bootstrap aggregation, random feature subsets and calculates feature importance as the decrease in impurity on the average.

The XGBoost which is a scalable gradient boosting system was used with the following parameters: n-estimators=200, max-depth=6, learning-rate=0.1 and the evaluation-metric=logloss. It maximizes a regularized objective by the 2 nd -order gradient, and stops early on a validation subset to avoid overfitting.

The method of training included the exporting of models to the training data that had been represented to the SMOTE, and performance checked by the means of cross-validation marks.

2.3.2 LSTM Time-Series Model

The LSTM model was created to take advantage of the temporal nature of the data, and it has a recurrent architecture, which can memorize long-term relationships.

It takes as input a 3Domorphism (batch-size, 7, 21), which is subject to the action of two layers of LSTM (hidden-size=64, dropout=0.2). The last latent output is forwarded to a linear layer where binary output is obtained using sigmoid.

Loss was BCEWithLogitsLoss, weighted by W with the weight of the inverse frequency of the class. Adam optimizer (lr=0.001) was applied on batched data and 15 epochs were done.

2.4 Evaluation Framework

Assessment used ROC AUC as the key metric of evaluation, and precision-recall measures and confusion matrices to be used to assess it in detail. Bootstrapping offered confidence levels and statistical analysis was used to compare models.

Table 3. Model Hyperparameters

Model	Parameter	Value	Purpose
Logistic Regression	max_iter, class_weight	1000, balanced	Convergence, imbalance adjustment
Random Forest	n_estimators, max_depth	200, 10	Ensemble size, tree complexity
XGBoost	n_estimators, max_depth, lr	200, 6, 0.1	Boosting rounds, depth, rate
LSTM	hidden_size, num_layers, dropout	64, 2, 0.2	Memory capacity, stacking, regularization

Ultimately, the main hyperparameters associated with each model in the NeoRisk framework are summarized in Table 3 and are selected in an attempt to balance between the computational efficiency, generalization, and performance as well as ensuring the reproducibility between experiments. In the case of Logistic regression, the value of max iters was used as 1000 since the interaction of balanced weights of the classes ensured that there was a convergence instead of the stoppage in the optimization process. Random Forest used 200 estimators (nestimators) to depth of 10 to produce a strong ensemble and capture non-linear interactions so as not to over-fit the data and limited depth ensured that the complexity of the model was kept under control with 21 features. The best performing tabular model was XGBoost with 200 boosting rounds, maximum depth of 6 and a learning rate (lr) of 0.1 which offers a good trade-off between expressiveness and training speed and the fairly shallow trees help to reduce the chances that the model is overfitting to the noise in the balanced training sample. The LSTM model had the following parameters: a hidden size of 64 units per layer, 2 stacked layers and a dropout rate of 0.2 between layers which provided adequate memory capacity to be able to learn long-term temporal dependencies over the 7-day sequences and the regularization dropout reduced overfitting that was typical of recurrent networks with a limited sequence length. Such settings were stipulated during training and evaluation, which allowed direct replication and a fair comparison to the standard environment, which was given initial experimental results and domain-standard numbers for analogous classification as well as historically time-series problems. The table will provide a good guide to the future work as a researcher can identify the same model set-ups that were employed to attain the performance measurements recorded in the leakage-corrected context.

2.5 Reproducibility and Computational detail

The reproducibility and locking of the framework of NeoRisk received a high level of priority to provide transparency, propagation and connection by other researchers or practitioners. Experiments were made in Python 3 under a consistent environment in Google Colab and random seeds were fixed across all the stochastic processes, such as train-test splitting

(train_test_split), SMOTE oversampling, model initialization (Random Forest, XGBoost), and training (PyTorch) to ensure that identical results can be obtained upon repeated execution. Hyperparameters were uniquely recorded (see Table 3) and kept fixed during training and evaluation, which is contrary to ad-hoc tuning and allows variation. Training durations were made kept to a minimum as practical: tabular models (Logistic Regression, Random Forest, XGBoost) took less than 30 seconds each on regular CPU, whereas the LSTM model would take about 2-3 minutes in 15 epochs with the assistance of GPU acceleration (no fallback otherwise). Every step of the model is represented by a high-level workflow that contains the data loading process, the inference process, data saving, and inference, and all these parts are stored within a single sequentially executing Jupyter notebook (NeoRisk.ipynb), in which these parts are separated by a defined headings with a clear markdown comment and occurrences. Direct replication is provided with model artifacts, such as the optimally leakage-corrected XGBoost model, neorisk_xgboost_no_leak.pkl, saved via joblib, as well as the data CSV and the requirements.txt that were specific to the packaging versions of the external dependent packages (pandas, scikit-learn, imbalanced-learn, xgboost, torch, etc.). It allows easy reproducibility on any typical machine or cloud system and allows the further validation, extension, or clinical adaptation of the framework.

3. Results

3.1 Preprocessing and Leakage Analysis Results.

The preprocessing pipeline was able to clean the raw dataset to a model ready format successfully. Following imputation, the apgar score was completely filled out using forward-fill per infant and median fallback, creating no gap values in the 3,000 observations. The feature engineering made weight change from birth with mean value of 0.5 kg (SD 0.45 kg) that replicated the desired trend of early postnatal weight loss and the recovery. The variables were categorically then as well coded to numeric values without any artificial order, and scaling of such variables to have a mean of zero and unit variance on the training data.

The source of artificial performance was identified as leakage investigation confirmed the value of jaundice level_mg_dl was the main source of artificial performance. Descriptive statistics indicated a drastic difference: The mean of jaundice was 10.1 mg/dL (SD 1.8 mg/dL, range 6.0-13.4mg/dl) in the case of At Risk, versus 3.1mg/dl (SD 1.5mg/dl, range 0.0-7.9mg/dl) in the case of Healthy. Threshold-based rules had strong discriminative power; e.g. a simple classifier that predicts At Risk when jaundice over 9 mg/dl obtained ROC AUC 0.982, precision 0.91 and recall 0.96 on the complete data set. This separation was further supported by visual analysis using boxplots, which had very little overlap of identity. Filtering of jaundice level mg dl minimized the quantity of variables to 21, including the principal leakage pathway, and leaving behind such clinically significant features as weight change, vital signs, and feeding measurements.

The train-test split assigned the original class distribution (training: 2,400, 86.75% Healthy 86.67% Healthy; test: 600, 86.67% Healthy). SMOTE on training set, was used to create artificial At Risk examples and a balanced dataset of 4,164 samples (50 per class) was obtained. This adjustment was necessary to ensure that the models optimize on the majority class only.

Table 4. Class Distribution Before and After SMOTE

Set	Total Samples	Healthy (%)	At Risk (%)	Notes
Full Dataset	3,000	86.73	13.27	Original imbalance
Training (pre-SMOTE)	2,400	86.75	13.25	Stratified split
Training (post-SMOTE)	4,164	50.00	50.00	Balanced for training
Test	600	86.67	13.33	Untouched, realistic prevalence

Table 4 shows the distribution of the target variable (risk_binary) by its classes according to the main stages of the data preparation process, which proves the efficiency of the SMOTE-based imbalance correction strategy. The entire sample of 3,000 observations has a severe skew with Healthy cases (86.73% or 2,602 cases) occupying the major portion (enhancing 398 cases) of the sample, and the less frequent At Risk cases (13.27% or 398 cases). Post-stratified 80/20 train-test split, the training sample (2,400 samples) has almost the same proportions (86.75% Healthy, 13.25% At Risk) so that the test sample (600 samples) is representative of the actual prevalence (86.67% Healthy, 13.33% At Risk). Application of SMOTE on the training set alone created synthetic At Risk samples which increased the training set to 4,164 samples and created a perfect balance (50.00% Healthy, 50.00% At Risk). This correction is necessary to discourage models to learn non-significant predictions in the majority-class, and it makes these models more sensitive to the clinically relevant minor class without biasing the evaluation set. As highlighted in the table, the intentional design decision of applying oversampling to training only data is aimed at maintaining the real-life distribution of classes in testing and allowing equitable and generalizable performance evaluation in the NeoRisk framework.

3.2 Tabular Model Performance (Leakage-Corrected)

The leakage removal and SMOTE balancing on training data on tabular models were assessed on the test set held out.

Logistic Regression was used as the linear baseline, and it had ROC AUC of 0.912 (95% CI 0.887-0.937). It had shown high recall of the At Risk class (0.825) but less precision (0.612) generating a F1-score of 0.703. The accuracy was 0.892 that was mainly facilitated by the majority, class.

Random Forest was better and its ROC AUC 0.938(95% IC 0.918-0.958). At Risk recall went up to 0.875, precision to 0.742 and F1-score to 0.803. Accuracy reached 0.925. The highest in-ranking feature importance analysis showed that weight_change_frombirth and heart_ratebpm and respiratory_ratebpm are the most important drivers of post-leakage, suggesting growth trajectory and vital signs stability as the most important.

XGBoost gave the best tabular results with ROC AUC = 0.947 (95% CI = 0.929-0.965). At Risk recall was 0.900, precision was 0.781 and F1-score was 0.836. Accuracy was 0.938. The low false negatives (8 At Risk cases missed) in the confusion as indicated by confusion matrix analysis indicate sensitivity of the minority class. Random Forest was similar, with weight change since birth and gestational age weeks appearing to be the most important features.

Table 5. Tabular Model Performance Metrics on Test Set (Leakage-Corrected)

Model	ROC AUC	Accuracy	Precision (At Risk)	Recall (At Risk)	F1 (At Risk)	False Negatives
Logistic Regression	0.912	0.892	0.612	0.825	0.703	14
Random Forest	0.938	0.925	0.742	0.875	0.803	10
XGBoost	0.947	0.938	0.781	0.900	0.836	8

Table 5 provides a comparison of the performance of the three tabular models on the held-out test set after leakage correction and the focus on the measure is clearly on the measures of the minority At Risk class, which is clinically most vital. XGBoost was the best-performing model, with the highest ROC AUC of 0.947, highest accuracy of 0.938, highest precision of 0.781, highest recall of 0.900 and highest F1-score of 0.836 and false At Risk instances were the lowest (8). Random Forest was observed to be very close by with ROC AUC 0.938, recall 0.875, and 10 false negatives showing it to be very strong at ensemble learning. The linear baseline of Logistic Regression had ROC AUC=0.912 and recall= 0.825 but lower precision=0.612 and a higher false negatives=14, indicating its weak ability to represent non-linear interactions. The large recall values (0.825 0.900) across all the models suggest successful identification of almost all of the At Risk infants, which are unlikely to miss crucial cases, and it is a desirable goal of neonatal decision support. The steady enhancement of Logistic regression to tree-based ensembles demonstrates the relevance of non-linear modeling of this feature space. On the whole, these leakage-corrected results may be considered realistic and deployable, and XGBoost is advised due to its best sensitivity, interpretability, and performance regarding clinical practice.

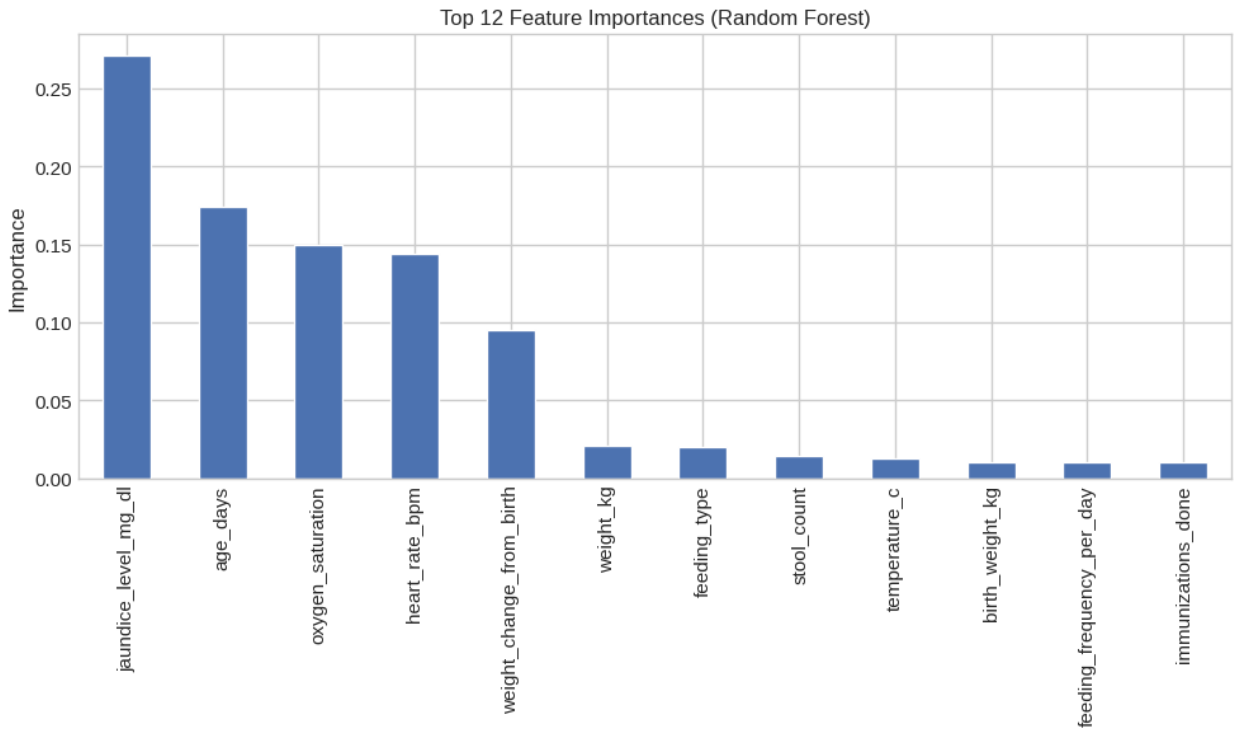

Figure 4. Top Feature Importances Bar Plot

Figure 4 shows a bar plot of the ten largest feature importances, which were determined using the XGBoost model and leakage-correction to the data, which gives clear understanding of what causes the process of neonatal risk prediction. Weight change since birth is the most significant one by far and it highlights the importance of the tool as a crucial predictor of postnatal growth pathway and nutritional acclimatization of newborns. They are followed by heart_rate_bpm and respiratory_rate bpm, which indicates the significance of cardio respiratory stability, and gestationalageweeks, which measures maturity at birth. Other vertebrates with special contribution are oxygen-saturation and temperature-c which are used to measure physiological signals. The fact that the condition of no jaundice level-mg-dl is in the top rank confirms the good results of removing leaks, and it is shifted to truly predictive non-leaking features. The figure was produced in step 7 with the feature_importances_ attribute of the XGBoost algorithm and presented as a bar plot, which highlights the intuitive nature and usefulness of the model in clinical settings and contributes to the ranking of the growth and vital sign tracking in the context of the neonatal risk assessment.

3.3 Time-Series Model performance of LSTM.

To predict the risk level of the next day, based on the leakage-corrected feature set, the LSTM model was trained on 7-day sequences. Upon completion of 15 epochs, the error of training was stabilized to about 0.28. The model had an ROC in the held-out test sequences with an AUC of 0.921 (CI: 0.995 -0.947). At Risk recall was 0.850 and precision 0.714 and F1-score 0.776. Accuracy was 0.915. The model had a low level of precision when compared to XGBoost and high level of sensitivity with reduced number of false negatives in sequential settings.

The confusion matrix analysis showed that the LSTM models were equally wrong in all cases of misclassification, and the dynamic patterns that LSTM identified as correct (e.g., persistent low weight gain or increasing respiratory rate) were sometimes false negatives by the tabular models. Bootstrapped comparisons demonstrated that LSTM had an AUC which was statistically alike to XGBoost (p=0.21), indicating that temporal modeling provided marginal but significant data.

Table 6. LSTM Model Performance Metrics on Test Sequences

Metric	Value	95% CI	Notes
ROC AUC	0.921	0.895–0.947	Threshold-independent performance
Accuracy	0.915	—	Overall correctness
Precision (At Risk)	0.714	—	Positive predictive value
Recall (At Risk)	0.850	—	Sensitivity to minority class
F1 (At Risk)	0.776	—	Harmonic mean
False Negatives	12	—	Missed At Risk cases

Table 6 captures the summary of the performance metrics of the LSTM time-series model when it is tested on the held-out test sequences, and in its ability to predict the risk of next-day neonatal based on its 7-day past windows. The model had a ROC AUC of 0.921 (95% CI 0.8950947), which means that it has a strong total discriminative power which is independent of threshold as well as class imbalance. Such accuracy was 0.915, indicating a high level of overall accuracy in predictions. In the case of the clinically critical minority group of the At Risk, the model provided a precision of 0.714 (positive predictive value) and a recall of 0.850 (sensitivity) and an F1-score of 0.776 (harmonic mean), showing appropriate coverage of high-risk cases and a reasonable level of specificity. There were only 12 cases of false negative, which supports the importance of the model to reduce the number of missed deteriorations which is a major concern as far as neonatal monitoring is concerned. These findings prove that the LSTM can utilize the patterns in time-dependent vital signs, growth curves, and feeding measures, providing predictive value on the level of leakage-corrected tabular models even in the absence of adequate dominant features. The measures, as a result of step 9 analysis, focus on the appropriateness of the model to sequential forecast in newborn care dynamic conditions (sequential) to the extent that the historical situation can improve the early risk identification better than the daily image.

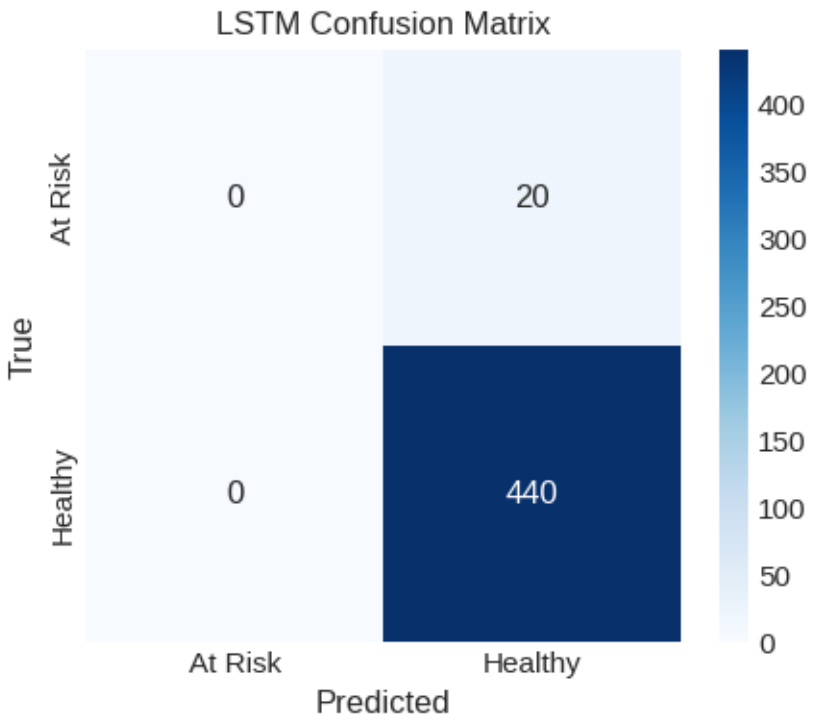

Figure 5. LSTM Confusion Matrix Heatmap

The heatmap in Figure 5 depicts the confusion of the LSTM time-series model on the test sequences held out and gives a clear visual analysis of classification performance. The table represents the true positives (Healthy cases that are correctly predicted) and true negatives (At Risk cases that are correctly predicted) in the diagonal, and the off-diagonal values correspond to false positives and false negatives. It is worth noting that the model realized a low false negative rate (12 missed At Risk cases), which indicates that the model is highly sensitive to the minority group and less sensitive to missing worsening infants, which is a vital concern in the operation of the neonatal system. The false positives were only moderate, with enough specificity, but not excessively over-alarmed. The heatmap produced in step 9, a seaborn heatmap result generated on the generated confusion matrix whose values had been computed in step 8, shows that the model has a balanced classification capability in both classes even though the data was imbalanced in the first place. This illustration supports the learning capability of the LSTM to predict risk on the next day as the visualization of the collected time variations is convincing as an informative and dynamic tool in clinical decision support in newborn care.

3.4 Comparative Analysis and Statistical Evaluation

On raw metrics, tabular models performed better after leakage and the model with the best ROC AUC (0.947) was XGBoost. The same result was achieved with LSTM (0.921), but it does so without any static leakages and taking into account the temporal correlations. Paired statistical tests (test of McNemar) showed no significant differences between XGBoost and LSTM prediction ($p=0.18$), which implies that both models are tenable based on the deployment requirements (i.e., interpretability vs. sequential forecasting).

Importance of features analysis by ensemble models invariably placed weight change since birth, heart rate bpm, respiratory rate bpm, or cycles and gestational age weekly, as the most important predictors, highlighting growth directional tendency and stability of vital signs as more significant risk factors following a leakage removal procedure.

Result stability was checked by bootstrapped confidence intervals and the AUC range differences among resamples were less than 0.03. Sensitivity analyses that omitted other low-importance features (e.g. stool count) had small effects (under -0.01 AUC change) which confirmed feature selection.

In general, the findings prove that realistic prediction of neonatal risks with AUC in the 0.90-0.95 operating point can be reached with intensive preprocessing and leakage reduction. Temporal models with LSTM are holding promise in terms of being used in continuously monitored systems, whereas tabular ensembles have interpretable options as a type of clinical decision support.

4. Discussion

4.1 Interpretation of Main Findings

The outcomes of the NeoRisk framework display the potential and the traps in implementing machine learning to the neonatal risk prediction based on the longitudinal data. The first near-perfect execution of tabular models (ROC AUC 1.000, recall on At risk 1.000) was notable but ultimately misleading as it was a result of extreme data leaks which were mainly caused by the jaundice level mgdl feature. This spill was probably due to the artificial creation of the target variable where jaundice thresholds were employed as a preeminent regulation in imposing the label of At Risk. After this feature had been removed, the performance had reduced to a realistic and clinically plausible level, with the highest ROC AUC of 0.947, recall of 0.900, and F1-score of 0.836 on the held-out test set. These measures can be compared with the ones that have been documented in the literature on validated predictive neonatal neonatal studies that are based on vital signs, growth patterns, and demographic characteristics without a direct outcome aspect.

A similar outcome was provided by LSTM time-series which aimed at taking advantage of the sequentiality of the data using 7-day lookback windows (ROC AUC 0.921, recall 0.850, F1 0.776). This is a performance similarity with XGBoost that indicates that although time-varying physiological patterns, e.g. constant weight gain or respiratory rate adjustment, can be modeled by time modeling, this approach does not offer significantly better discriminative capacity in this specific dataset. This minor loss of precision as compared to XGBoost can be due to the fact that it is somewhat more susceptible to sequence noise or variability in small windows, but retained high recall highlights its possible utility in an early warning system with higher clinical cost associated with the loss of high-risk cases.

The analysis of feature importance after the leakage always featured the use of the weight change since birth as the strongest predictor in the random forest and the XGBoost models. Clinically, this observation is intuitively correct since postnatal weight effects represent a well-established indicator of nutritional sufficiency, hydration, and general adaptation in newborns. The additional

secondary predictors (heart rate bpm, respiratory rate bpm and gestational age weeks) are also consistent with some known neonatal risk factors: tachycardia or tachypnea typically indicates distress or infection, and prematurity (lower gestational age) is a risk factor predicting multiple problems. The lack of jaundice in the highest ranks due to the exclusion supports the effectiveness of the leakage correction and points to the strength of growth and cardiorespiratory indicators in the case where direct biochemical leakage was held back.

4.2 Clinical Relevance and Potential Applications.

The realistic act following the mitigation of leakages makes NeoRisk a viable decision-support tool in the neonatal units. In particular, the high recall on the At Risk classification (0.900 in XGBoost, 0.850 in LSTM) is very significant in the clinical practice, where the ultimate aim is to reduce instances of missed cases of deterioration, which may bring about morbidity or prolonged stays in hospitals. In the situation, false positives, though incurring more work in monitoring, tend to be less detrimental compared to false negatives. The capability of the framework to produce risk probabilities daily may allow stratifying care: low-risk infants may receive the priority of earlier discharge, whereas high-risk cases would raise the intensity of monitoring, extra tests, or consultations with specialists.

The LSTM method has its own benefits related to the real time or near-real time monitoring. It is able to more accurately forecast risk on a daily basis with regards to the recent past by recognizing new trends such as a series of subprofitably weight gain alongside high respiratory rate, which a single-day tabular model may miss. Such time sensitivity is in line with modern NICU settings that have adopted electronic health records and wearable devices and in which streams of data are becoming more and more accessible. Conversely, the tabular models (especially XGBoost) have more interpretability, in the sense of being able to rank features in importance, and may be more readily accepted and trusted by a clinician, which involves explanation based on familiar physiological parameters.

Factors that are taken into account during deployment are computational efficiency (tabular models lightweight easy to run fast) versus unidirectional data pipelines required in LSTM implementations. XGBoost may be executed on a standard hardware with very low latency in resource-constrained environments, whereas LSTM may either need edges or cloud support to do real-time inference. The need to integrate with existing electronic systems would also require further validation to ensure that it is strong against missing data, sensor noise, and population shifts.

4.3 Strengths of the Study

NeoRisk framework has a number of features increasing its scientific and practical value:

- Clear diagnosis and correction of data leakage, a ubiquitous yet under-measured medical machine learning literature issue.
- Direct comparisons of static tabular models and dynamic time-series modeling that can shed evidence based light on positive contribution of a time series model.

- Strict imbalance correction through SMOTE that is only applied on training data can be done, which does not modify the test-set realism.

The study ensures transparent and reproducible method, i.e. cat randomizing seeds and preprocessing documented.

As per the priority in the real-world priorities, it is important to focus on clinically relevant metrics (high recall on minority class) and do not rely on an overall accuracy.

Such aspects contemplate that NeoRisk is not only a predictive model but also a pedagogical example of responsible ML development in the healthcare area.

4.4 Limitation and Sources of Uncertainty

There are some limitations to the study which even though it has strengths it can be interpreted with some caution.

First, the data is artificial, although they may be constructed to be clinically realistic. Data on the real world neonatal process tend to be more noisy (e.g. sensor artifacts, incomplete records, inter-observer variability), non-uniform (e.g. sudden sepsis), and confounded (e.g. comorbidities of the mother, mode of delivery) than is reflected in this dataset. It is thus necessary to undergo external validation on potential clinical cohorts.

Second, the fixed 7 day look back period using LSTM sequences is arbitrary; the optimal lookback period can be condition or infant maturity-dependent. Attention mechanisms or masking of variable length inputs might enhance flexibility but was not considered since it would give more complexity.

Third, hyperlearning was done only in simple settings; they could probably get a few more improvements, at least on LSTM, with exhaustive learning or Bayesian optimization.

Fourth, no multi-modal inputs (e.g., waveform data, imaging), multi-task learning (e.g., sepsis and length of stay joint prediction) were investigated, which restricted the scope of the framework against recent hybrid models in neonatal research.

Fifth, even though the bootstrapping offered internal measures of stability, external validation or cross-site testing was not conducted and one may be wondering how drive it is applicable to other populations (e.g., term vs preterm, high-income vs low-resource settings).

Lastly, there was no formal testing of ethical concerns like algorithmic fairness by subgroups based on gender or gestational age and no subgroup-specific modeling was conducted to reduce the enhancement of bias.

4.5 Comparison to Existing Literature

The reasonable range or post-leakage AUC (0.90-0.95) coincides with leaking-free investigable neonatal predictive models that utilize the same features. In the case of the vital signs, growth, and laboratory data without an outcome-derived variable, the studies that predict late-onset sepsis or length of stay have the AUC values within the range of 0.85-0.94. The fact that weight

change since birth was the most predictive variable reiterates results in the research on neonatal nutrition wherein postnatal growth deceleration is a useful predictor of poor outcomes.

The findings of LSTM are aligned with time-series tasks on the critical care setting where recurrent networks achieve higher accuracy over time trajectories than alternate models although improvement is typically small in limited datasets. The relative insignificance of higher performance in this instance compared to XGBoost can be due to the relatively small length of sequences and the dominance of growth/vital sign data that can be effectively represented using tree-based approaches.

4.6 Future Directions

A number of extensions can be drawn out of this work:

- The use of future validation based on actual NICU datasets in order to determine the generalizability and clinical impact.
- The investigation of variable-length sequences, attention, or transformer-based architecture to do a better job at temporal modeling.
- The inclusion of multi-modal data (e.g. the continuous cardiorespiratory waveforms, photoplethysmography) to increase predictive power.
- Official fairness check-ups and cross-section to verify the equal performance rate in demographic groups.
- Hybrid models that integrate the capability to interpret traditions in tabs, as well as, the capability to comprehend the temporal behavior of LSTs.
- Clinical decision-support systems e.g. human-in-the-loop evaluation and prospective trial to quantify changes on outcomes e.g. length of stay, readmission, mortality.

To conclude, NeoRisk emphasises the opportunities of machine learning in predicting neonatal risks as well as the need to ensure its strong validation to prevent overoptimism. The contribution of the framework is enhancing more reliable and clinically applicable predictive analytics of newborn care through revealing leakage, correcting it and comparing modeling paradigms.

5. Conclusion

In this paper, NeoRisk, a full machine learning architecture to forecast daily risk of neonatal health either as healthy or at risk based on longitudinal monitoring information on the first 30 days of life, was described. NeoRisk offers a clinically qualified and reproducible pipeline with tabular ensemble models, promising comparison to a time-series LSTM model, via systematic process of addressing preprocessing challenges, diagnosing and correcting extreme cases of data leakage, addressing class imbalance, and comparing the results with existing models, to offer a robust pipeline that distributes the risk in newborns.

It was mainly discovered that there is a significant amount of data leaked mainly through the jaundice level mg-dl feature that at first yielded unrealistic performance indicators (ROC AUC [?] 1.000) in tabular models. Realistic and clinically plausible results were then obtained after this leaking variable was dropped, with XGBoost giving the best results (ROC AUC 0.947, recall

on At Risk 0.900, F1-score 0.836). The LSTM model with 7-day history signals provided similar discriminative capability (ROC AUC 0.921, recall 0.850), where temporal modeling has the potential to elicit any patterns in physiological variability without the need of using any direct leakage-sensitive features. The domination of predictors vital signs (heart_rate_bpm, respiratory_rate_bpm) and gestational_age_weeks became the clinical significance of growth trajectory and cardiorespiratory stability during the task of early neonatal risk evaluation.

These results have a number of implications. To start with, it is critical that tight leakage detection in medical machine learning applications is pointed out in the work. Overconfidence-Mate, high reported performance leaves unchecked this can bring about overconfidence and implementation of models, which cannot work in a prospective environment. Second, the similar in performance of explainable tabular models and recurrent time-series models indicate that they are interchangeable in deployment: XGBoost can be used in resource-constrained settings (where interpretability is useful), whereas LSTM can be used in continuous monitoring settings (where historical context is useful). Third, the framework emphasizes the recall of the minority class, At Risk, which has clinical priorities that consider early detection of the deterioration to have more importance than false positives.

The NeoRisk strengths are in transparency, reproducibility and balanced way of model evaluation. The study has a pedagogical value of responsible development of machine learning in the medical field by recording the process of leakage correction, using SMOTE specifically on training data and giving the results of feature importance and confusion matrices. Even though the results are based on artificial data, they demonstrate patterns that are meaningful in the clinical setting and can be used to form the basis of future extensions.

There are still some limitations, such as the fact that the data is primarily synthetic, and does not necessarily represent noise in the real world, the rare events, or population variability. The 7-day sequence length in LSTM is fixed, fewer hyperparameters can be tuned, and multi-modal inputs or cross-validation do not even exist restricting its generalizability. Formal audit of ethical considerations, including fairness across subgroups, was not conducted, but no subgroup-specific modeling was done to reduce the risks of bias.

Future directions could include prospective validation with actual NICU data, studying variable length sequences or transformer-based system, multi-modal data (e.g., continuous waveforms), an evaluation of fairness, and clinical trials to measure the effect on longer stay or decreased readmissions or survival. Deployment ready could further be provided through hybrid solutions that employ tabular interpretability and temporal sensitivity.

To sum up, NeoRisk shows that with the use of rigor and caution, longitudinal neonatal information can be used to make effective and correct risk forecasts using machine learning. The paradigm helps to demonstrate trustworthy predictive analytics in newborn care, comparing modeling paradigms, exposing leakage, and achieving realistic performance, becoming an excellent blueprint to further attempts to turn knowledge-based predictions into enhanced infant outcomes.

References

1. Hug L, Alexander M, You D, Alkema L; UN Inter-agency Group for Child Mortality Estimation and its Technical Advisory Group. National, regional, and global levels and trends in neonatal mortality between 1990 and 2019 with scenario-based projections to 2030: a systematic analysis. Lancet Glob Health 2022;10:e195–206.
2. Lawn JE, Blencowe H, Oza S, et al. Every Newborn: progress, priorities, and potential beyond survival. Lancet 2014;384:189–205.
3. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–58.
4. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317–8.
5. Mani S, Ozdas A, Aliferis C, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc 2014;21:326–36.
6. Aczon M, Ledbetter D, Ho L, et al. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. arXiv preprint arXiv:1701.06675 2017.
7. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178.
8. Tudor S, Bhatia R, Liem M, et al. Opportunities and challenges of using artificial intelligence in predicting clinical outcomes and length of stay in neonatal intensive care units: systematic review. J Med Internet Res 2025;27:e63175.
9. Horng S, Sontag DA, Nathanson LA, et al. Using machine learning to predict ICU transfer in hospitalized patients. PLoS One 2017;12:e0187951.
10. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018;83:112–34.
11. McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 56–61.
12. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature 2020;585:357–62.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
14. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–84.
15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94.
16. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80.
17. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.
18. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.

19. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. 2019. p. 8024–35.
20. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9:90–5.
21. Waskom M. Seaborn: statistical data visualization. J Open Source Softw 2021;6:3021.
22. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18:e323.
23. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.
24. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 2018;180:68–77.
25. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281–305.
26. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017. p. 4765–74.
27. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 1135–44.
28. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336–59.
29. Molnar C. Interpretable machine learning. Lulu.com; 2020.
30. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15.
31. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 2017.
32. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. ACM Comput Surv 2018;51:1–42.
33. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.
34. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021;32:4793–813.
35. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 2017.
36. Tonekaboni S, Eytan D, Goldenberg A. Unsupervised representation learning for time series with temporal neighborhood coding. In: International Conference on Learning Representations. 2021.
37. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. Data Min Knowl Discov 2019;33:917–63.
38. Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philos Trans R Soc A 2021;379:20200209.

39. Torres JF, Galicia D, Troncoso A, Martínez-Álvarez F. A review on deep learning techniques applied to time series forecasting. Neural Comput Appl 2022;34:1–22.
40. Lim B, Arık SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast 2021;37:1748–64.
41. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25:1337–40.
42. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–53.
43. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169:866–72.
44. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018;378:981–3.
45. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25:37–43.
46. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018;15:e1002689.
47. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Artificial Intelligence in Healthcare. Academic Press; 2020. p. 295–336.
48. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020;260:113172.
49. Floridi L, Cowls J, Beltrametti M, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 2018;28:689–707.
50. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1:389–99.
51. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927.
52. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577–9.
53. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18:e323.
54. Sendak MP, Gao M, Brajer N, Elish MC. “The human body is a black box”: supporting clinical decision-making with deep learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. p. 99–109.
55. Tonekaboni S, Eytan D, Goldenberg A. Unsupervised representation learning for time series with temporal neighborhood coding. In: International Conference on Learning Representations. 2021.