Development And Validation Of A Prediction Model For Mortality In Children Aged Under Five Years With Clinical Pneumonia In Rural Gambia

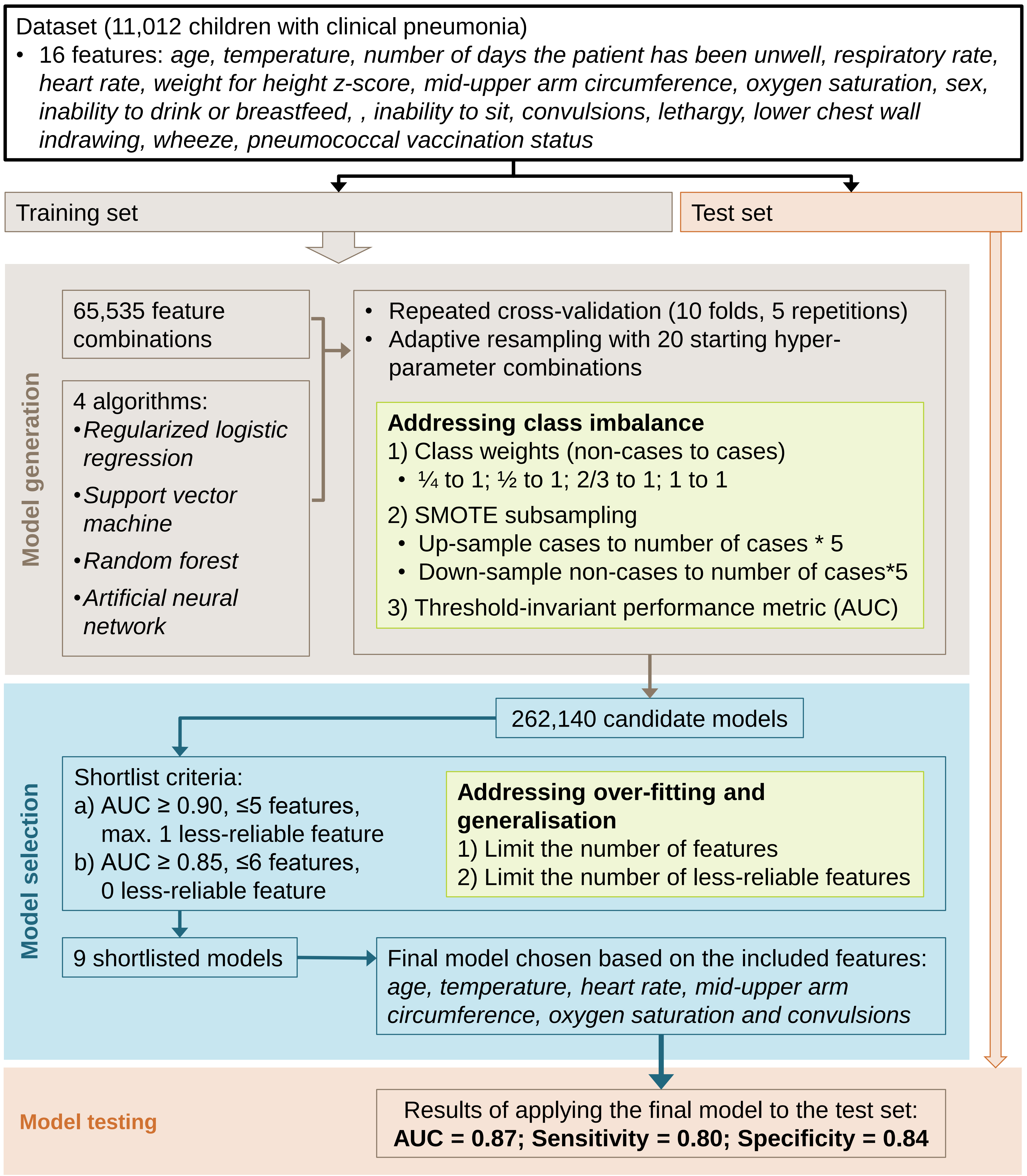
Alexander Jarde[1], David Jeffries[1], Grant Mackenzie[2]

[1]Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine, Department of Statistics and Bioinformatics
[2]Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine, Disease Control and Elimination

Introduction

Pneumonia accounts for a high percentage of deaths in children under the age of five in developing countries. A reliable and generalizable tool to predict mortality and thus assess the severity of pneumonia would aid patient management.

Our goal was to use machine learning algorithms to build a predictive model that performed well not only on the training set, but that also on new, unseen data, increasing the likelihood that it would generalize to other datasets.



Dataset (11,012 children with clinical pneumonia)
- 16 features: *age, temperature, number of days the patient has been unwell, respiratory rate, heart rate, weight for height z-score, mid-upper arm circumference, oxygen saturation, sex, inability to drink or breastfeed, , inability to sit, convulsions, lethargy, lower chest wall indrawing, wheeze, pneumococcal vaccination status*

Training set

Test set

Model generation

65,535 feature combinations

4 algorithms:
- *Regularized logistic regression*
- *Support vector machine*
- *Random forest*
- *Artificial neural network*

- Repeated cross-validation (10 folds, 5 repetitions)
- Adaptive resampling with 20 starting hyper-parameter combinations

Addressing class imbalance
1) Class weights (non-cases to cases)
   - ¼ to 1; ½ to 1; 2/3 to 1; 1 to 1
2) SMOTE subsampling
   - Up-sample cases to number of cases * 5
   - Down-sample non-cases to number of cases*5
3) Threshold-invariant performance metric (AUC)

Model selection

262,140 candidate models

Shortlist criteria:
a) AUC ≥ 0.90, ≤5 features, max. 1 less-reliable feature
b) AUC ≥ 0.85, ≤6 features, 0 less-reliable feature

Addressing over-fitting and generalisation
1) Limit the number of features
2) Limit the number of less-reliable features

9 shortlisted models

Final model chosen based on the included features: *age, temperature, heart rate, mid-upper arm circumference, oxygen saturation and convulsions*

Model testing

Results of applying the final model to the test set:
AUC = 0.87; Sensitivity = 0.80; Specificity = 0.84

Figure 1: *Flow diagram of the development, selection and testing of a predictive model for mortality in infants with clinical pneumonia.*

Methods

Datasets: We used a dataset of 11,012 children admitted with clinical pneumonia, with data on 16 features (variables) and on each children's survival. The dataset was split into 2 subsets based on the date of admission: one to develop the prediction model (7341 subjects, 2/3 of the dataset) and a test set to evaluate its predictive performance in new data (3671 subjects, 1/3 of the dataset).

Model generation: For each possible combination of two or more features we used four machine learning algorithms to generate predictive models: support vector machine, neural networks, random forests and regularized logistic regression. Each model was developed using repeated cross-validation (5 repetitions, 10 folds) with adaptive resampling to optimize the tuning hyper-parameters. To address the challenge of having a very imbalanced data set (only 2% of deaths) we

1) applied each algorithm using four different class weighing schemes (penalising misclassifications of deaths more or less),
2) used the Synthetic Minority Over-Sampling Technique (SMOTE) to balance the number of events, and
3) used a threshold-invariant metric (Area Under the ROC Curve) to assess a model's performance.

We used a high performance computer to generate a model for each of the 65,535 feature combinations with each of the four algorithms.

Model selection: With the assistance of two clinicians, we classified each feature as reliable or less reliable to identify those whose measurement would be more homogeneous across populations. In order to increase the likelihood that the chosen model would generalize well (no overfitting) we shortlisted those that not only had an excellent performance, but that also used a limited number of features and a limited number of less reliable features. The final model was chosen among the shortlisted ones based on what particular features it included.

Model testing: We tested how the final model performed on new, unseen data by applying it to the test set.

Results

The final model included age, temperature, heart rate, mid-upper arm circumference, oxygen saturation and convulsions. Not only did it have excellent sensitivity and specificity (both >85%) on the training set, but more importantly, it had promising performance when applied to the test set, with sensitivity = 0.80 and specificity = 0.84 (AUC = 0.87).

Conclusion

Our predictive model performed well not only in cross-validated data, but also in our test dataset, increasing our confidence in its generalizability.

Acknowledgements

We would like to thank the assistance of Dr. Lobga Galega Babila and Dr. Yekin Ajaoi Olatunji for their assistance rating the features as less-reliable or not.