A validation of machine learning models for the identification of critically ill children presenting to the paediatric emergency room of a tertiary hospital in South Africa: A proof of concept

M A Pienaar,[1] PhD; **N Luwes,**[2] DTech; **J B Sempa,**[3] PhD; **E George,**[4] PhD; **S C Brown,**[5] MD

[1] *Paediatric Critical Care, Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa*
[2] *Central University of Technology, Bloemfontein, South Africa*
[3] *Department of Biostatistics, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa*
[4] *MRC Clinical Trials Unit, University College London, UK*
[5] *Paediatric Cardiology, Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa*

***Corresponding author**: M A Pienaar (PienaarMA1@ufs.ac.za)*

Background. Machine learning (ML) refers to computational algorithms designed to learn from patterns in data to provide insights or predictions related to that data.
Objective. Multiple studies report the development of predictive models for triage or identification of critically ill children. In this study, we validate machine learning models developed in South Africa for the identification of critically ill children presenting to a tertiary hospital.
Results. The validation sample comprised 267 patients. The event rate for the study outcome was 0.12. All models demonstrated good discrimination but weak calibration. Artificial neural network 1 (ANN1) had the highest area under the receiver operating characteristic curve (AUROC) with a value of 0.84. ANN2 had the highest area under the precision-recall curve (AUPRC) with a value of 0.65. Decision curve analysis demonstrated that all models were superior to standard strategies of treating all patients or treating no patients at a proposed threshold probability of 10%. Confidence intervals for model performance overlapped considerably. Post hoc model explanations demonstrated that models were logically coherent with clinical knowledge.
Conclusions. Internal validation of the predictive models correlated with model performance in the development study. The models were able to discriminate between critically ill children and non-critically ill children; however, the superiority of one model over the others could not be demonstrated in this study. Therefore, models such as these still require further refinement and external validation before implementation in clinical practice. Indeed, successful implementation of machine learning in practice within the South African setting will require the development of regulatory and infrastructural frameworks in conjunction with the adoption of alternative approaches to electronic data capture, such as the use of mobile devices.
Keywords. Machine learning, critical care, children, domain knowledge, triage, severity of illness (min. 5 - max. 8).

South Afr J Crit Care 2024;40(3):e1398. https://doi.org/10.7196/SAJCC.2024.v40i3.1398

Contribution of the study
This study presents the first application of machine learning to identify critically children in South Africa. The use of machine learning models in the critical care environment has the potential to improve decision making and improve patient outcomes.

Measures to improve the care of critically ill children are an important concern in low- and middle-income countries (LMICs).[1] In South Africa (SA), Hodkinson *et al.*[2] have demonstrated that failures in the identification, resuscitation and referral of critically ill children contribute to avoidable escalations of severity of illness and mortality in SA. Promoting early recognition and escalation of life-saving care, consultation with expert clinicians, and vigilance and referral of these patients may improve the outcomes of critically ill children in SA.

Machine learning (ML) refers to computational algorithms designed to learn from patterns in data to provide insights or predictions related to those data.[3] Publications reporting the use of ML in triage appear in the international literature.[4-6] To date, however, no such models have been described in SA or other LMICs. Statistical models developed for the prediction of paediatric intensive care unit (PICU) mortality in developed countries have shown variable performance in LMICs, with the authors suggesting that variations in patient characteristics and

case mix contribute to these variations.[7] These data support a need to develop such models in LMIC settings and, together with the rapidly increasing interest in and use of this technology, underpin the rationale for the research presented here.

In previous work, our group investigated the use of an artificial neural network (ANN) model for mortality prediction in two South African PICUs.[8] Expanding on this research, we developed nine ML models in a single SA centre for the identification of critically ill children.[9] We made use of ANNs,[10] extreme gradient boosting (XGB)[11] and logistic regression (LR).[10,11] The models aim to predict the presence of paediatric critical illness as a composite outcome of death before hospital discharge or admission to the PICU. This study presents a prospective, internal validation of these novel ML models in an LMIC setting, using previously unseen data from the same centre. The validation provides an assessment of model discrimination and calibration, decision curve analysis, and model explanation.

Methodology

Study site

This study was conducted at a regional referral hospital in SA. Patients were enrolled in the paediatric emergency oom. This is the primary site of referral to the specialist paediatrics service for children requiring evaluation for acute illness. Patients requiring admission to the hospital are transferred to the general paediatric wards or the tertiary PICU within the same hospital. Patients requiring other subspecialty care are referred to another facility in the same area from the general paediatric ward or PICU. Admission to PICU occurs on a consultation basis where severity of illness, likelihood of clinical success, provision of specific life-supporting care and availability of resources are considered.

Study population and sampling

Validation data were collected prospectively from 1 March to 30 June 2022. Patients under 13 completed years of age presenting with acute illnesses (duration <7 days) were enrolled in the study. Patients presenting dead on arrival or for scheduled clinic visits or elective procedures were excluded from the study. While no specific guideline exists for determining the ideal size of a test or validation dataset,[12] a proposed split of 75:25 between training (756 patients) and test data was selected, and a minimum validation sample size of 252 participants was determined from the size of the training data.[9,12]

Test data were collected by treating clinicians at the point of care using the same measurements and observations ordinarily made during patient assessment. A mobile device data collection platform was used to collect data directly into a REDCap database. The details of the data collection procedure are provided in Table 1.

Study outcome

The outcome variable in this study was a composite outcome of death before hospital discharge or admission to the PICU, which served as a surrogate for the presence of critical illness in these children.

Statistical analysis

Descriptive analysis was conducted using Python 3 in the Jupyter Notebooks environment. Categorical data were presented using frequencies and percentages, while continuous data were presented using means with standard deviations (SDs), medians with interquartile ranges (IQRs), and 95% confidence intervals (CIs).[14,15]

Models

We evaluated nine candidate ML models that had been previously developed. These included three ANNs, three extreme gradient boosting (XGB) and three logistic regression (LR) models. Three sets of variables were selected in development and one of each model type was trained for each set. Training data for these models included 756 participants. These models are summarised in Table 2. We have reported on model development and hyperparameter tuning elsewhere.[9]

Model performance was analysed in terms of discrimination, calibration and decision curve analysis (DCAs). Discrimination was determined by construction of receiver operating characteristic (ROCs) and precision recall (PRCs) curves with their respective areas under the curve (AUROCs, AUPRCs). Given the tendency of ROC analysis to be overoptimistic when imbalance exists between classes,[16] PRC analysis was employed as the main parameter of discrimination to ensure a more realistic assessment. The threshold for a random classifier was considered to be the event rate (0.12) and the minimum possible AUPRC calculated using the formula from Boyd *et al.*[17] was 0.06.While no clear standard for AUPRC exists, we considered higher AUPRC preferable as this best represents the trade-off between false negative classifications and false positive classifications in imbalanced data sets. For the assessment of AUROC, values >0.7 are considered acceptable, >0.8 excellent and >0.9 exceptional. The threshold for a random classifier was set at 0.5.[18]

Calibration was determined by the degree to which predicted probabilities from models agreed with real probabilities of an outcome.[19] We considered the lowest level of calibration to be the agreement between mean predicted probability and the event rate. Flexible calibration curves were constructed for all models. Models where the slope of the flexible calibration curve was close to 1 and the intercept was close to 0 were considered weakly calibrated. Moderately calibrated models were considered to meet the above criteria and be close to the ideal calibration line from [0,0] to [1,1]. Strong calibration refers to the idealistic goal of near-perfect calibration of predictions to event rates in all categories of prediction.[20] This approach provides a more robust assessment of calibration over a range of probabilities than the widely reported Hosmer-Lemeshow statistic.[21]

We conducted DCA to determine the net benefit of models. This approach includes aspects of discrimination and calibration and determines the net benefit of a model compared with baseline strategies of intervening in all or no patients.

$$\text{Net benefit} = \text{sensitivity} \times \text{prevalence} - (1 - \text{specificity}) \times (1 - \text{prevalence}) \times w$$

where w is the odds at the threshold probability. We compared models in terms of net benefit models and the number of interventions avoided across the range of thresholds compared with an approach that intervenes in all patients. The reader is directed to the guide by Vickers *et al.*[22] for a more comprehensive description of this method. We considered models with higher DCA curves above the x-axis and greater numbers of avoided interventions superior. We did not consider threshold probabilities >50% in DCA as this would assign greater importance to false positive classifications. In this analysis, we suggest a 10% predicted probability as a possible threshold for assigning an intervention, accepting that this threshold could differ between providers and institutions.

Where clinicians implement ML or predictive models in practice, it is necessary that, in addition to the performance metrics above, they understand how or why models make certain abovementioned

Table 1. Data collection procedure

Variable	Measurement	Recorded value
Age		Continuous (months)
Level of consciousness	Subjective clinician assessment: Alert – age-appropriate level of awareness and interaction. Equivalent to AVPU scale of A. Not alert – any decreased level of consciousness.	Binary
Unable to feed	Parental report of inability to achieve adequate oral feeding during this illness.	Binary
Respiratory distress	Subjective clinician assessment: tachypnoea, subcostal or intercostal recessions, or nasal flaring.	Binary
Weak pulses	Subjective clinician assessment of radial pulse character.	Binary
Respiratory rate	Manual clinician count >one minute.	Continuous (breaths per minute)
Capillary refill time	The time to return of colour after five seconds of finger pressure to the sternum in infants or finger pulp in children.	Continuous (seconds)
SPO^2	Mindray VS9 vital signs monitor.	Continuous (%)
SBP and DBP		Continuous: Mean blood pressure = DBP + 0.333 (SBP - DBP). Then converted to *z*-score for age (13) mmHg
Pulse rate (from SPO^2)		Continuous
Capillary blood glucose	StatStrip Xpress 2 glucometer	Continuous

SPO^2 = saturation of oxygen by pulse oximetry; SBP = systolic blood pressure; AVPU (a scale of consciousness): A = alert, V = response to verbal stimuli, P = response to pain, U = unresponsive; DBP = diastolic blood pressure.

Table 2. Candidate models

Model	Features included in model
ANN1	Respiratory rate *z*-score
XGB1	Peripheral oxygen saturation
LR1	Pulse rate
	Mean blood pressure *z*-score
	Capillary refill time
	Quantitative hypoglycaemia (value <3 mmol/L)
	Quantitative hyperglycaemia (value >10 mmol/L)
	Respiratory distress (the presence of recessions, nasal flaring, grunting or head bobbing)
	Weak pulses
	Level of consciousness (alert or not alert)
	Inability to feed (inability to consume adequate fluids or food by the oral route)
ANN2	Mean blood pressure zscore
XGB2	Quantitative hypoglycaemia
LR2	Quantitative hyperglycaemia
	Respiratory distress
	Level of consciousness
	Inability to feed
ANN3	Respiratory rate
XGB3	Peripheral oxygen saturation
LR3	Quantitative hypoglycaemia
	Quantitative hyperglycaemia
	Level of consciousness
	Age
	Inability to feed

ANN = artificial neural network; XGB = extreme gradient boosting; LR = logistic regression.
*Three groups of features were selected for model development and one of each algorithm type was trained on each group, leading to a total of 9 models.

recommendations.[23,24] To achieve this, we made use of SHapely Additive exPlanations (SHAP) to generate post hoc explanatory representations of model predictions.[25] In the bee swarm plots presented, binary values are represented as pink for 1 and blue for 0 and continuous variables are represented as pink for higher and blue for lower. Values that increase the predicted probability from the model appear further to the right on the x-axis; for example, if lower pulse oximetry values (blue) appear further to the right on the x-axis. We determined that models where predictions were logically coherent with clinical knowledge would be preferred.

Ethical clearance

Internal review board approval was obtained from the Health Sciences Research Ethics Committee (ref. no. UFS-HSD2020/2204/2505-0003) and the Free State Department of Health (ref. no. FS_202102_019). Informed consent was obtained in writing from the legal guardians of children participating in the study. Assent was obtained from children capable of doing so. All data were stored on a secure server and were fully anonymised before exporting as a CSV file for analysis.

Results

Descriptive analysis

The validation study included 267 participants. This provided a train:test split of 74:26 overall. There were three deaths (1.1%) and 32 PICU admissions (12.0%). The composite outcome was noted in a total of 33 patients (12.4%). One participant died without admission to the PICU. Table 3 presents the descriptive analysis of the data.

Model performance

The ROC and PR curves of the three best-performing models (highest AUPRC) are presented in Fig. 1. All models are presented in Table 4 and Supplementary Fig. 1 (https://www.samedical.org/file/2305).

Most models had an AUROC of at least 0.8, indicating excellent discrimination.[19] The highest AUROC score recorded was 0.84,

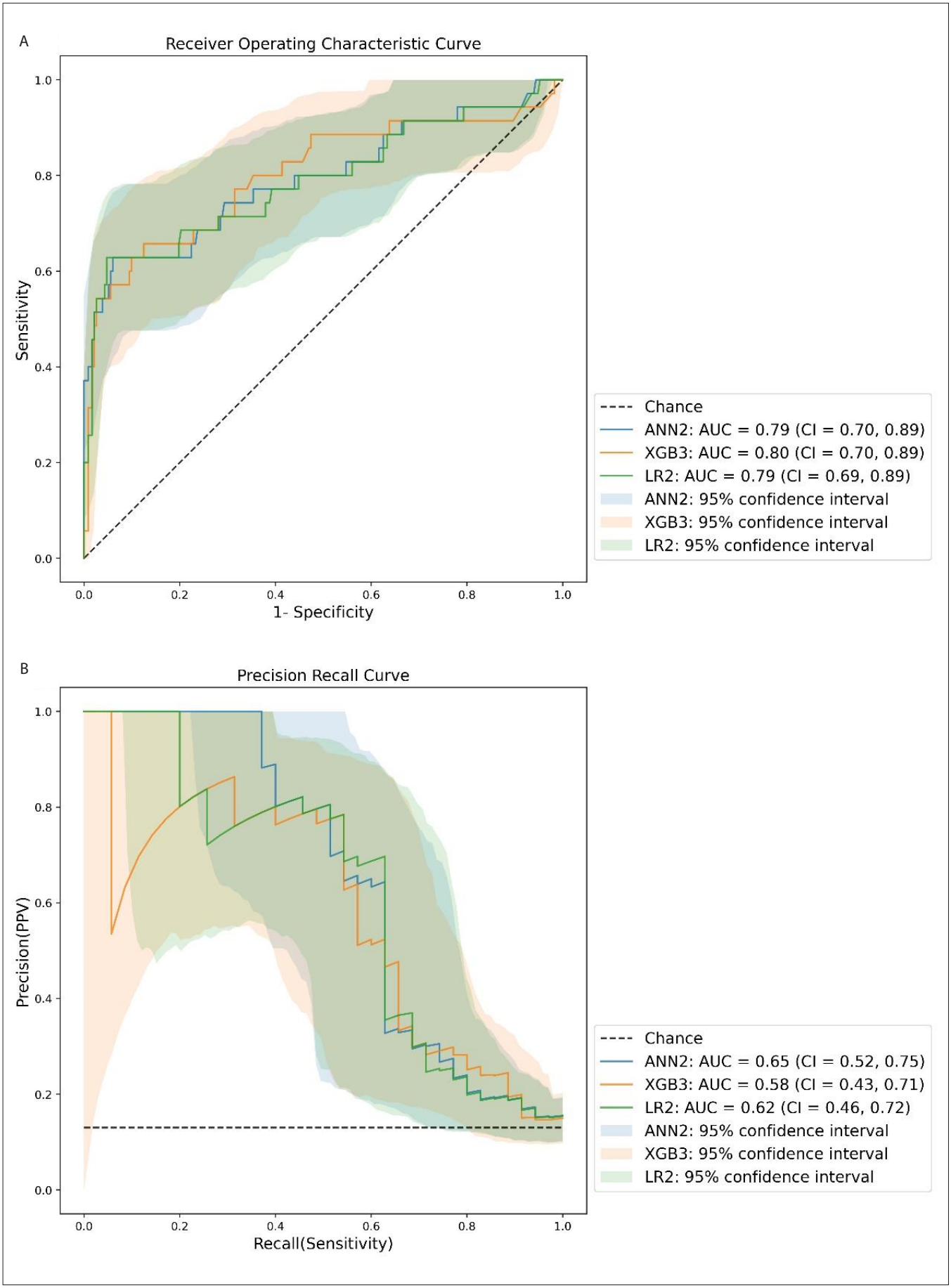

Fig. 1. Receiver operating characteristic (A) and precision recall (B) curves.

achieved by ANN1. All models demonstrated an AUPRC greater than the event rate (0.12) with a minimum of 0.5 in ANN3. The highest AUPRCs were 0.65 in ANN2, 0.62 in LR2 and 0.58 in XGB3. Model-wide metrics are summarised in Table 4. Flexible calibration curves could only be constructed with four bins owing to empty bins, limiting their utility. Most calibration slopes included 1 in their CIs, and most intercepts 0, but model calibration is likely weak at best.

All models demonstrated greater net benefit compared with treat all or treat none strategies. All models were able to reduce the number of interventions per 100 patients compared with a treat-all strategy (Fig. 2).

Model interpretation

The aggregated importance of features across the ML models is presented as a bee swarm plot in Fig. 3.

In the bee swarm plot, the value for each variable is represented on a colour scale (blue being low or negative and pink being high or positive). The relative impact of each figure on model prediction is provided as a SHAP value on the x-axis. Thus, from this figure, models predicted higher probabilities of the outcome where, for example, if a patient was found to not be alert, or have a lower pulse oximetry reading, the predicted probability of the study outcome would be higher. The SHAP values presented above provide an easily interpreted depiction of how features impact model predictions. The model-agnostic nature allows comparisons between different algorithms. The explanations presented are intuitive and coherent with clinical domain knowledge.

Table 3. Descriptive data analysis

Continuous features		Categorical features	Frequency (%)
Age (months)	Median = 10.0	Weak pulse	
	IQR = 3, 30	No	250 (93.6)
		Yes	17 (6.4)
Respiratory rate (breaths per minute)*	Mean = 41.8	Level of consciousness	
	SD = 14.3	Alert	225 (84.6)
		Decreased	42 (15.4)
Peripheral oxygen saturation (%)*	Median = 98.0	Unable to feed	
	IQR = 96, 100	No	235 (88.0)
		Yes	32 (12)
Pulse (beats per minute)	Median = 144.0	Respiratory distress	
	IQR = 125, 163	No	198 (74.2)
		Yes	69 (25.8)
Mean blood pressure (mmHg)†	Mean = 73.0		
	SD = 15.7		
Capillary refill time (seconds)	Median = 2.0		
	IQR = 2, 3		
Glucose (mmol/L)	Mean = 5.6		
	SD = 2.8		

IQR = interquartile range; SD = standard deviation.
*Missing = 1.
†Missing = 5.

Discussion

This study aimed to validate the performance of novel ML models designed to identify critically ill children presenting to a paediatric emergency unit at a single centre in SA. All models demonstrated the ability to identify patients at risk of the study composite outcome of death before hospital discharge or admission to the PICU. Model performance correlated well with performance metrics calculated in stratified cross-validation during the development phase.[9] The development data set was derived from autumn to summer and is thus trained on data that take into account seasonal variation in the patient case mix. The population was similar to the development data set, in which 3.92% of the study population died, 12.94% of the population required PICU admission, and 15.16% of the population experienced the combined outcome. The development population is described in the development study.[9]

Table 4. Model-wide performance metrics

	ANN1		ANN2		ANN3	
Metric	**Score**	**CI**	**Score**	**CI**	**Score**	**CI**
AUROC	0.84	(0.77, 0.87)	0.80	(0.75, 0.85)	0.76	(0.71, 0.81)
AUPRC	0.53	(0.45, 0.57)	0.63	(0.57, 0.69)	0.52	(0.45, 0.57)
ECE	0.05	(0.02, 0.08)	0.07	(0.04, 0.1)	0.06	(0.03, 0.09)
Slope	1.10	(0.8, 1.4)	1.15	(0.83, 1.47)	1.23	(0.83, 1.63)
Intercept	-0.02	(-0.04, -0.0)	-0.03	(-0.05, -0.01)	0.03	(0.01, 0.05)
	XGB1		**XGB2**		**XGB3**	
AUROC	0.80	(0.75, 0.85)	0.79	(0.74, 0.84)	0.80	(0.75, 0.85)
AUPRC	0.54	(0.46, 0.58)	0.56	(0.5, 0.62)	0.58	(0.52, 0.64)
ECE	0.05	(0.02, 0.08)	0.03	(0.01, 0.05)	0.04	(0.02, 0.06)
Slope	0.98	(0.76, 1.2)	1.03	(0.81, 1.25)	1.01	(0.77, 1.25)
Intercept	-0.04	(-0.06, -0.02)	-0.05	(-0.08, -0.02)	-0.02	(-0.04, -0.0)
	LR1		LR2		LR3	
AUROC	0.82	(0.77, 0.87)	0.79	(0.74, 0.84)	0.80	(0.75, 0.85)
AUPRC	0.57	(0.49, 0.61)	0.62	(0.56, 0.67)	0.55	(0.47, 0.59)
ECE	0.04	(0.02, 0.06)	0.04	(0.02, 0.06)	0.02	(0.00, 0.04)
Slope	1.13	(0.87, 1.39)	1.05	(0.81, 1.29)	1.15	(0.87, 1.43)
Intercept	-0.01	(-0.02, 0.0)	-0.05	(-0.08, -0.02)	-0.01	(-0.02, 0.0)

ANN = artificial neural network; CI = 95% confidence interval; AUROC = area under the receiver operating characteristic curve; AUPRC = area under the precision-recall curve; ECE = expected calibration error; XGB = extreme gradient boosting; LR = logistic regression.

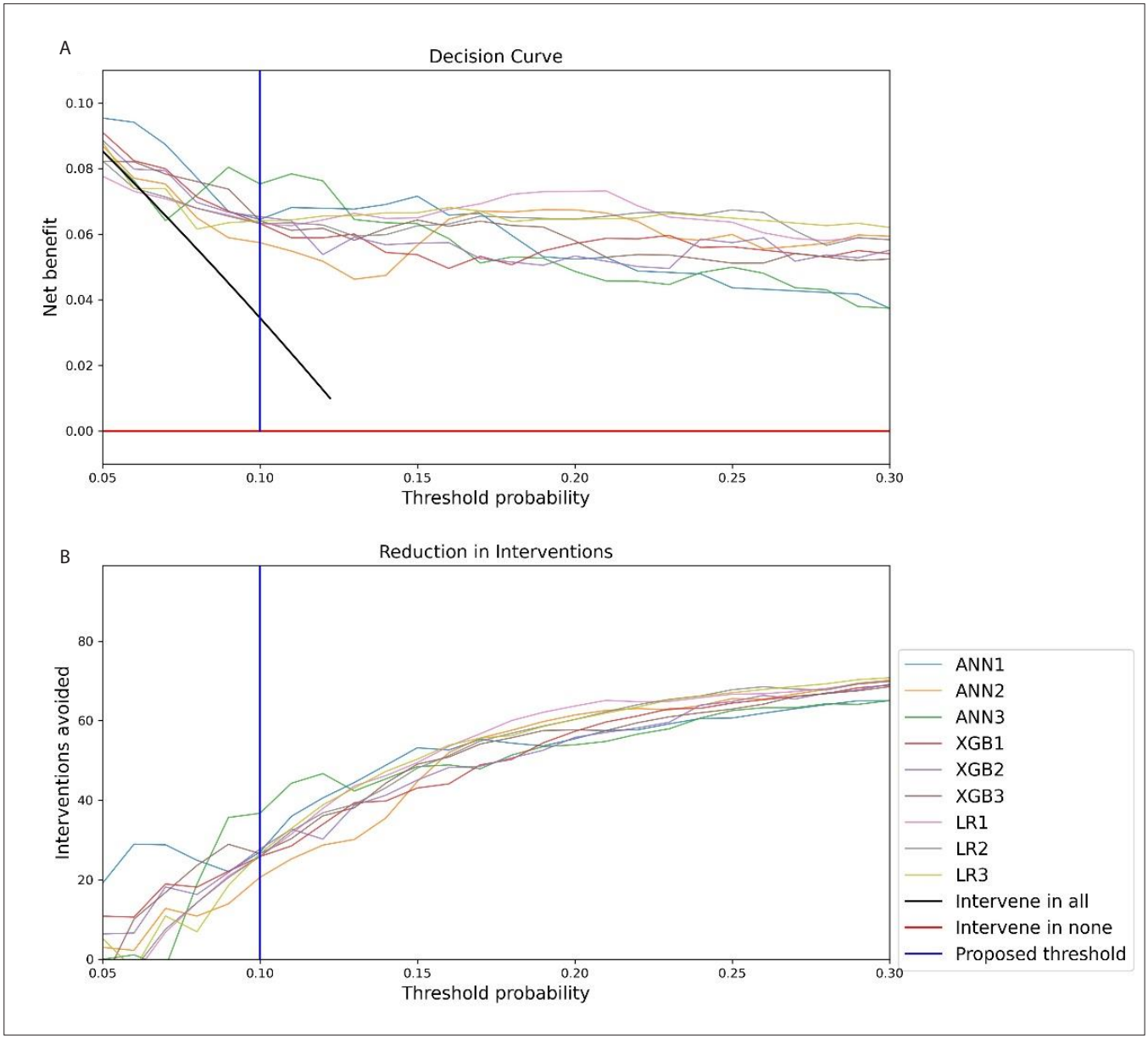

Fig. 2. Decision curve analysis.

Model performance was reported as comprehensively as possible, and the reporting of single metrics of performance was avoided. This provides a transparent account of the strengths and weaknesses of the models assessed.[26] The developed models demonstrated good discrimination and could successfully differentiate between the two research classes (children who died or were admitted to the PICU, and children who survived without PICU admission) at a given threshold. As such, a proof of concept is established for these models from a modelling perspective. Models demonstrated weak calibration, in terms of the framework of Van Calster *et al.*[20] That is to say, models did not systemically overestimate or underestimate risk and were not overly or underly confident. However, calibration is less relevant to this application, as the goal in future applications is to discriminate between children who are critically ill or not, as opposed to providing a highly accurate probability. Calibrated probabilities would be more relevant in other applications such as estimating prognosis to communicate to families.

Evaluation of models relevant to their envisioned application is critical to ensuring that models provide maximal benefit in the clinical setting. In this study, models are intended to trigger appropriate responses directed towards preventing avoidable escalations of severity of illness, ultimately with the aim of reducing mortality and morbidity. While true negative and true positive predictions are unlikely to incur costs in this study, false negative and false positive predictions may incur significant costs.[27] False negative classifications may expose patients to a significant risk of harm if they delay activation of appropriate responses as well as potentially increasing healthcare workload and expenditure due to increased severity of illness. Similarly, false positive classifications may incur costs such as increased healthcare worker workload, avoidable healthcare expenditure and exposure to potentially harmful and unnecessary interventions and investigations. While the exact value of these costs could not be determined, we propose that false negative classifications are more costly in this context. Decision curve analysis demonstrated that the developed models offered benefit compared with either a treat all or treat none approach and offered the ability to reduce interventions compared with a treat all approach.

While these models demonstrated the ability to discriminate between critically and non-critically ill children, the CIs for model performance metrics were wide and preference for one model or another could not be

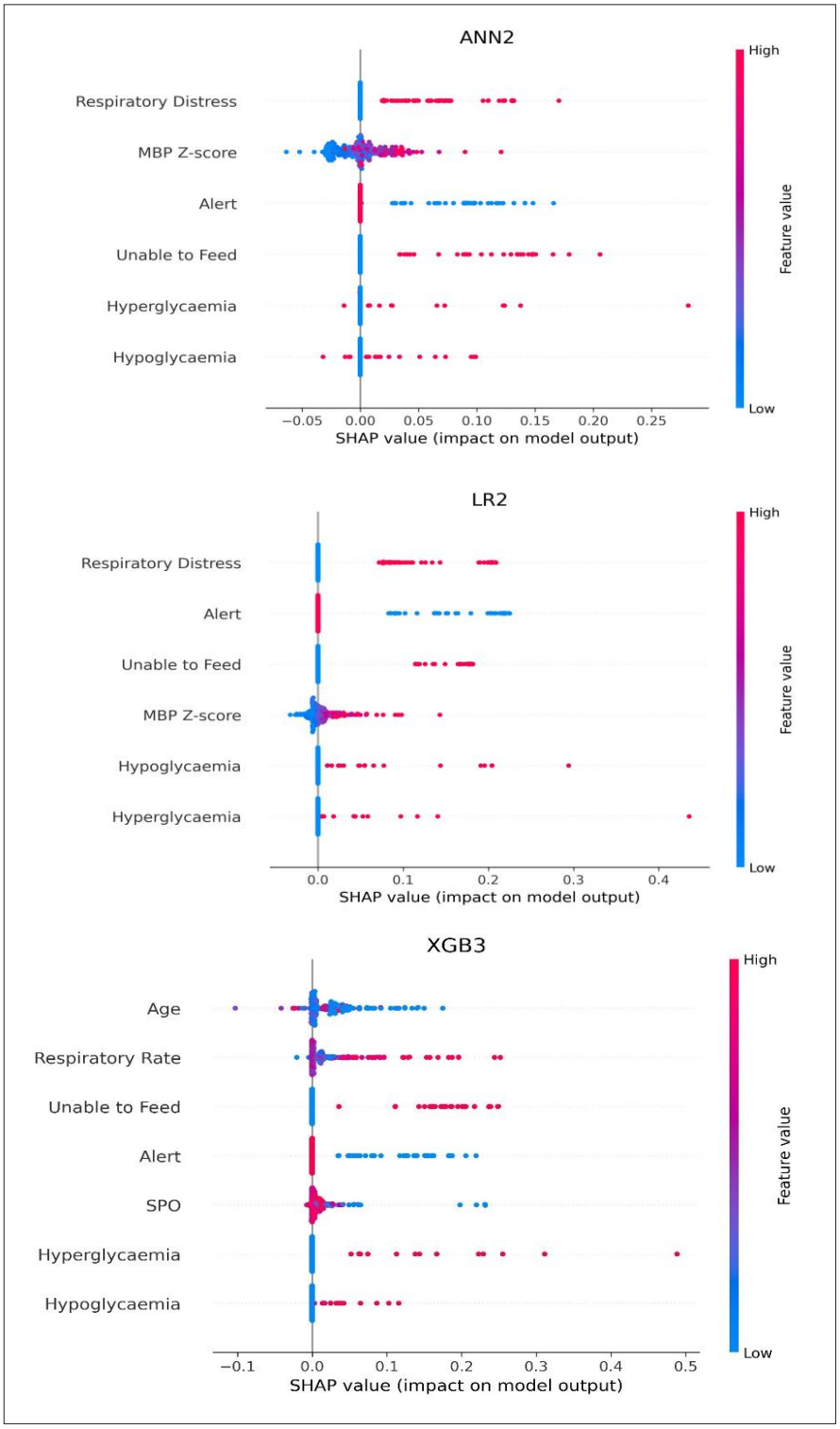

Fig. 3. SHAP feature importances. (SPO = peripheral pulse oximetry measurement.)

established. As such, they cannot be employed in clinical practice in their current iterations. It is worth pointing out that even with refinement and perfect model performance, these models will never be intended for autonomous use. They will remain a clinical tool to be used by trained professionals to facilitate improved care of critically ill children by enabling earlier identification of serious illness.

Post hoc model explanations generated with SHAP found that model predictions were intuitive and coherent with clinical reasoning. This makes the models likely to be acceptable for use by clinicians in practice, particularly if they are provided with model explanations together with predicted probabilities.

The models presented here differ from those reported by Goto *et al.*,[4] Gaita *et al.*[5] and Hwang and Lee.[6] These studies had access to large electronic datasets in their development. Goto *et al.*[4] included 52 037 participants in the United States in their study while Hwang and Lee[6] included 2 621 710 participants from a large national database in South Korea. This difference in data availability is a cardinal difference between these studies and the research presented. Lack of access to useable electronic health records (EHR) has been identified as a challenge in SA[22] and addressing this limitation is a core recommendation from the findings of this study. Goto *et al.* achieved an AUROC of 0.85 for the prediction of PICU admission or death during hospitalisation, using a deep neural network, superior to the best performing model in this study. Hwang and Lee developed a random forest model for the prediction of critical illness (in their study defined as admission to PICU, resuscitation in the emergency department, or death in the emergency department). The developed model achieved a AUROC of 0.99 and AUPRC of 0.64. Of note, in our study, the AUPRC of ANN2 of 0.62 is comparable to that of Hwang and Lee.

The findings of this internal validation study are thought to be a favourable proof of concept and suggest that a larger, multi-centre refinement and validation study would be warranted as would retraining and updating of models, including a larger corpus of multi-centre data from SA. It may also underpin progress to studies of implementation or clinical efficacy.

While a digital healthcare policy[28] exists in SA, this document does not directly address a pathway to implementation of applied ML in clinical practice. As this study demonstrates the proof of concept that such models can be developed in the SA setting, it is important that policy follow such progress. Relevant domains in this goal would include standards for reasonable validation, the development of pipelines and infrastructure for data storage and use, and the development of robust ethical, regulatory and legal frameworks for implementation of ML in research and practice.

Study strengths

To our knowledge, this is first SA study of its kind. All models were developed using simple, readily available clinical features to potentially maximise future utility across a wide spectrum of healthcare environments. In contrast to previous studies undertaken in high-income

countries, this study utilised mobile devices to capture clinical data, thereby overcoming the lack of established electronic healthcare records in local healthcare facilities. Thus, the study has successfully presented a novel adaptation of ML methods to resource constraints. Furthermore, this study serves as an important proof of concept and offers an important methodological reference for further studies in this field.

Study limitations

This study was subject to some limitations, the most significant arguably being the relatively small sample size, which is reflected in the wide confidence intervals for model performance metrics. As an internal validation, the metrics reported only reflect the performance of models in this single centre, thus limiting generalisability. The need for informed consent prior to enrolment may have limited the inclusion of some unstable patients, possibly explaining the relatively low overall case mortality rate. The developed models were not compared with current triage practices, preventing comparison of benefit between these models and conventional approaches.

Clinical implications and relevance

With further refinement and investigation, these models may improve the identification of critically ill children by integration within SA's proposed digital health framework.[28] By implementing these models on a smartphone or tablet, even healthcare providers with limited skill or experience may trigger appropriate responses such as involvement of senior personnel, resuscitation or referral to higher levels of care or critical care environments.

Recommendations

External validation and studies of implementation and clinical utility are required to further the findings in this study. While model performance is promising, implementation in practice is not yet feasible. The evaluated models require considerable further refinement before implementation. These models should be externally validated and updated with a significantly larger data set in a multi-centre validation study in SA. Further research is also required to evaluate the best clinical use of such models and benchmark them against current practice.

Conclusion

In this study, we present the internal validation of ML models for the identification of critically ill hospitalised children in South Africa on a small sample of previously unseen data. To our knowledge, this is the first report of this kind. Models were evaluated in terms of discrimination, calibration, decision curve analysis and SHAP analysis. All models demonstrated satisfactory discriminatory performance in internal validation, correlating well with the findings in cross-validation from the development study. The superiority of one model could not be demonstrated. Nonetheless, overall performance is promising in terms of potential benefit in clinical practice. Model explanations demonstrated that model predictions are in keeping with clinical knowledge, an important finding when considering clinical implementation. Further refinement and external validation in a range of settings are required before investigating implementation in clinical practice. A modern, pragmatic approach to the collection of electronic health data underpinned by a robust regulatory framework are essential steps in the pursuit of effective ML applications in future clinical practice.

Declaration. None.

Acknowledgements. The authors thank the clinicians who aided in collection of data.

Author contributions. MAP conceptualised and designed the study, oversaw data collection, conducted data preparation, and designed and tested the ML models. NL acted as PhD co-supervisor, contributed to conceptualisation and design, advised during ML modelling and provided critical inputs in manuscript preparation. JBS contributed to conceptualisation and design, provided inputs into statistical analysis and provided critical inputs in manuscript preparation. EG contributed to conceptualisation and design, provided inputs into statistical analysis and provided critical inputs in manuscript preparation. SCB acted as PhD supervisor in the study, contributed to conceptualisation and design and provided critical inputs in manuscript preparation.

Data availability. The datasets generated and analysed during the current study are available from the corresponding author on reasonable request. Any restrictions or additional information regarding data access can be discussed with the corresponding author.

Funding. This research was funded by the National Research Foundation through the Thuthuka PhD Grant (TTK200412512685).

Conflicts of interest. None.

1. Turner EL, Nielsen KR, Jamal SM, von Saint A, Musa NL. A review of pediatric critical care in resource-limited settings: A look at past, present, and future directions. Front Pediatr 2016;4(Feb):1-15.
2. Hodkinson P, Argent A, Wallis L, et al. Pathways to care for critically ill or injured children: A cohort study from first presentation to healthcare services through to admission to intensive care or death. PLoS One 2016;11(1):1-17.
3. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14): 1347-1358.
4. Goto T, Camargo Jr CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. JAMA Netw Open 2019;4:2(1):e186937.
5. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019;23(1):1-13.
6. Hwang S, Lee B. Machine learning-based prediction of critical illness in children visiting the emergency department. PLoS One 2022;17(2): e0264184
7. Solomon LJ, Naidoo KD, Appel I, et al. Pediatric index of mortality 3-an evaluation of function among ICUs in South Africa. Pediatr Crit Care Med 2021;22(9):813-821.
8. Pienaar MA, Sempa JB, Luwes N, Solomon LJ. An artificial neural network model for pediatric mortality prediction in two tertiary pediatric intensive care units in South Africa. A development study. Front Pediatr 2022;10.
9. Pienaar MA, Sempa JB, Luwes N, George EC, Brown SC. Development of artificial neural network models for paediatric critical illness in South Africa. Front Pediatr 2022;10.
10. Graupe D. Principles of Artificial Neural Networks. 3rd ed. Singapore: World Scientific; 2013.
11. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (accessed 27 July 2022). https://doi.org/10.1145/2939672.2939785
12. Splitting a dataset into train and test sets. Baeldung on computer science (accessed 28 April 2022). https://www.baeldung.com/cs/train-test-datasets-ratio
13. Roberts JS, Yanay O, Barry D. Age-based percentiles of measured mean arterial pressure in pediatric patients in a hospital setting. Pediatric Crit Care Med 2020;E759-768.
14. Van Rossum GDF. Python 3 Reference Manual. Scotts Valley: CreateSpace; 2009.
15. Kluyver T, Ragam-Kelley B, Perez F, Granger B. Jupyter notebooks - a publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas. Amsterdam: IOS Press.
16. 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(3):1-21.
17. Boyd K, Costa VS, Davis J, Page CD. Unachievable region in precision-recall space and its effect on empirical evaluation. Proceedings of the 29th International Conference on Machine Learning, ICML 2012. 2012;1:639-646.
18. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thoracic Oncol 2010;5(9):1315-1316.
19. Schmid CH, Griffith JL. Multivariate classification rules: Calibration and discrimination. In: Encyclopedia of Biostatistics. Hoboken: John Wiley & Sons; 2005.
20. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167-176.
21. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. BMC Med 2019;17(1).
22. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3(1):1-8. https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0064-7
23. Holzinger A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform 2016;3(2):119-131.

24. Holzinger A. From machine learning to explainable AI 2018. https://hci-kdd.org
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 201722; 2017-December:4766–75. https://arxiv.org/abs/1705.07874v2 (accessed 6 May 2022).
26. 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(1):55-63.
27. Elkan C. The Foundations of Cost-Sensitive Learning. San Diego: University of California; 2001.
28. National Department of Health. National digital health strategy for South Africa 2019 - 2024. Pretoria: National Department of Health. 2019. pp 1-36. http://www.health.gov.za/wp-content/uploads/2020/11/national-digital-strategy-for-south-africa-2019-2024-b.pdf

Received 24 August 2024; accepted 3 October 2024.