



Hybridized neural networks for non-invasive and continuous mortality risk assessment in neonates

Stephanie Baker^{a,*}, Wei Xiang^b, Ian Atkinson^c

^a College of Science & Engineering, James Cook University, Cairns, Queensland, 4878, Australia

^b School of Engineering and Mathematical Sciences, La Trobe University, Melbourne, Victoria, 3086, Australia

^c Research Centre, James Cook University, Townsville, Queensland, 4811, Australia

ARTICLE INFO

Keywords:

Machine learning
Neural networks
Neonatal mortality
Mortality risk prediction
Prognostics
Intensive care

ABSTRACT

Premature birth is the primary risk factor in neonatal deaths, with the majority of extremely premature babies cared for in neonatal intensive care units (NICUs). Mortality risk prediction in this setting can greatly improve patient outcomes and resource utilization. However, existing schemes often require laborious medical testing and calculation, and are typically only calculated once at admission. In this work, we propose a shallow hybrid neural network for the prediction of mortality risk in 3-day, 7-day, and 14-day risk windows using only birthweight, gestational age, sex, and heart rate (HR) and respiratory rate (RR) information from a 12-h window. As such, this scheme is capable of continuously updating mortality risk assessment, enabling analysis of health trends and responses to treatment. The highest performing scheme was the network that considered mortality risk within 3 days, with this scheme outperforming state-of-the-art works in the literature and achieving an area under the receiver-operator curve (AUROC) of 0.9336 with standard deviation of 0.0337 across 5 folds of cross-validation. As such, we conclude that our proposed scheme could readily be used for continuously-updating mortality risk prediction in NICU environments.

1. Introduction

Complications resulting from premature birth are the leading cause of death in children under 5 [1], and over 50% of neonatal deaths occur in preterm infants [2]. Child deaths due to preterm birth are in excess of 1.1 million per year globally [3]. Recent data shows that preterm birth rates are increasing in 62 of the 65 countries with reliable trend data, indicating that this is a growing problem throughout the world.

Preterm infants are regularly cared for in Neonatal Intensive Care Units (NICUs). A recent study in the United States found that 84.41% of very low birthweight infants (those weighing 500–1499 g) and 41.18% of low birthweight infants (those weighing 1500–2499 g) are admitted to NICU, respectively [4]. In the NICU, assessment of mortality risk assists medical specialists in making difficult decisions regarding which treatments should be used and when, and whether initiated treatments are working effectively. It has been identified that precise mortality prediction would ease the process of making such decisions [5].

Currently, there are several scoring schemes used in NICUs for mortality risk assessment. One commonly used score is the updated Clinical Risk Index for Babies (CRIB-II) [6], which is a recalibrated and

simplified iteration of the original CRIB score [7]. Another family of scores that are routinely used are the Score for Neonatal Acute Physiology (SNAP) [8] and its derivatives, which include the expanded SNAP Perinatal Expansion (SNAPPE) [9], and the simplified versions of SNAP-II and SNAPPE-II [10]. The Berlin score [11] and Neonatal Mortality Prognostic Index (NMPI) [12] are also used, albeit to a lesser extent.

There are several limitations with the existing scores. Firstly, all aforementioned scores include parameters that require complex manual measurement. Additionally, the scores were all developed over 15 years ago. A recent review of the existing scoring systems has identified the need for updated and enhanced scores based on more recent cohorts [13]. This is further highlighted by a recent study [14], which found that the SNAPPE-II score achieved a significantly lower area under the receiver-operator curve (AUROC) of 0.849 on babies admitted between 2012 and 2013 compared to the AUROC of 0.92 when SNAPPE-II was proposed in 2001 [10]. Similarly, a recent paper [5] found that CRIB-II achieved AUROCs of 0.667 and 0.708 for mortality cases in ≤ 7 days and > 7 days, respectively, on babies admitted between 2001 and 2011 - a drastic decrease from the AUROC of 0.92 reported in the 2003 paper that

* Corresponding author.

E-mail address: stephanie.baker@jcu.edu.au (S. Baker).

proposed CRIB-II [6].

With recent studies identifying the weaknesses of existing scores, there has been some renewed interest in developing updated neonatal mortality risk scores using new techniques, however this field is in its infancy. Several studies have investigated techniques including logistic regression [15,16], densely-connected neural networks [17], random forest [18], and fusion of multiple machine learning algorithms into a super learner [19]. Of these, the highest performing schemes in terms of AUROC were those that used neural network techniques [17,19] and random forest [18].

The literature on predicting adult mortality is far more extensive, and many studies have investigated machine learning for prediction of mortality in adult ICU [20–27], with most achieving reasonable ability to distinguish between mortality and non-mortality. Of particular interest are long short-term memory (LSTM) neural networks, which were identified to be suitable for mortality prediction and the related problem of sepsis prediction in several works [20,27,28]. One recent work [22] also investigated the combination of convolutional neural network (CNN) layers with LSTM layers, achieving an AUROC of 0.836 using 48 h of laboratory and vital sign data. In our own previous work, we developed a scheme for adult mortality risk assessment in intensive care using features based on basic demographics and variations in heart rate, respiratory rate, blood pressure, blood oxygen saturation and temperature as inputs to a CNN-LSTM neural network, with significant success [29]. Much of the adult mortality literature is relevant to neonatal mortality, as similar strategies and variables can be used to develop machine learning approaches for prediction of mortality on the neonatal cohort.

A recurring limitation in much of the literature for mortality prediction is the selection of variables that are tedious or difficult to measure regularly. Such parameters include extensive laboratory results [19], maternal characteristics [15,17], and existing conditions [17,19]. This limits the usefulness of such schemes, as often the acquisition of these parameters would increase the burden on neonatal healthcare staff. Conversely, several other studies were limited by their selection of variables that do not change - such as the scheme that considered pre-birth and start-of-labour characteristics [15] and the scheme that used birthweight, blood oxygen at admission, and respiratory support within the first 24 h from birth [16]. This prevents recalculation of the infant's risk on a continuous or ongoing basis, and does not allow for assessment of response to treatments.

An ideal mortality risk prediction scheme would be one that uses fundamental demographics and routinely measured parameters to provide continuous mortality risk assessment, allowing for assessment of changing risk throughout the NICU stay without placing unreasonable additional burden on NICU staff.

As such, this paper proposes the Neonatal Artificial Intelligence Mortality Score (NAIMS), a hybrid CNN-LSTM neural network that relies on simple demographics and trends in heart and respiratory rate to determine mortality risk in the NICU for short- and long-term risk windows. Using 12 h of data from any window, NAIMS shows strong performance in predicting an infant's risk of mortality within 3, 7, or 14 days. This is the first known work to consider only basic demographics along with respiratory rate and heart rate statistics to produce an accurate prediction of immediate mortality risk. This scheme would provide rapid assessment of mortality risk in several risk windows, with no need for invasive procedures or knowledge of medical histories. Due to the simplicity and high performance of our proposed scheme, NAIMS could readily be continuously and automatically recalculated, enabling analysis of a NICU baby's responsiveness to treatment and other health trends. To the best of the authors' knowledge, NAIMS is the first scheme to utilize a hybrid CNN-LSTM network to predict neonatal mortality from demographics and vital signs.

2. Methodology

2.1. Data selection

The data used in this study was obtained from the Medical Information Mart for Intensive Care (MIMIC-III) clinical database [30]. This database includes records from 7870 neonates admitted between 2001 and 2008. As this study focuses on all infants admitted to the NICU for any reason, the criterion used to select patients was that the first care unit was the NICU. No exclusions were made based on birthweight, gestational age, or any other factors.

At this stage, the length of stay for the mortality cases was evaluated to determine the most useful windows for mortality prediction. It was found that the average length of stay (LOS) was 8.08 days with a high standard deviation (SD) of 16.75 days across the overall patient cohort. However, LOS was lower amongst the mortality subset of the cohort, with a mean of 4.66 days and SD of 5.58 days. The full distribution of LOS within the database is illustrated by the boxplots in Fig. 1.

The distribution of LOS, clustered around the lower end, served as our motivation for considering several risk windows of varying lengths, namely 3-day, 7-day, and 14-day windows. These windows were also utilized in our previous work [29] with high success, further supporting the use of these windows in this work. Assessment of mortality risk within these three windows would enable assessment of immediate risk, as well as longer-term survival prospects. As this model can be recalculated regularly throughout the stay, it is also suitable for the outlier cases with extremely long LOS values.

Given our largest window of interest is 14 days, we acquired 14 days of data for each patient. If the NICU stay exceeded 14 days, the first 14 days were obtained for non-mortality cases, while the 14 days prior to death time were used for the mortality cases. Where any patient stay was less than 14 days, all data from NICU admission to discharge or death were obtained.

Information obtained from this database included gestational age, birthweight, sex, time of death (where applicable) and available chart events for respiratory rate (RR) and heart rate (HR). No further demographic, health, or other factors were considered, as our aim was to develop a tool that could be applied with minimal prior knowledge of the patient. Maternal characteristics were also not considered. Feature selection was then performed, as discussed in the following subsection.

2.2. Feature selection

In selecting features for our proposed NAIMS scheme, there were several major considerations. Firstly, to prevent placing additional burden on healthcare staff, we determined that features must be based on parameters that are easy to measure. Ideally, dynamic parameters would also be able to be measured automatically. Secondly, we selected features based on recent findings in the literature.

Demographics features that describe fundamental information about the patient were selected - namely birthweight, gestational age, and sex. Birthweight and gestational age have repeatedly been shown in the literature to be strong indicators of mortality risk, and have been used by most existing schemes in the literature for this reason. Birthweight is a static variable, and thus the first birthweight in the patient's record was used. For most patients, gestational age was recorded in MIMIC-III as a range (e.g. 26-28 weeks). As such, we took the mean of the provided age range to be the gestational age. Where the infant was older than 40 weeks, their gestational age was recorded as "40" in MIMIC-III; we thus took this age group as simply 40.

Sex has also been used as it has been long known that physiological differences between the sexes lead to differing normal ranges for vital signs [31]. In our preprocessing, we set babies' sex to '1' or '0', corresponding to the 'F' or 'M' classification in MIMIC-III, respectively.

Next, we selected commonly recorded parameters in the NICU. A recent comprehensive review paper [32] concluded that the current

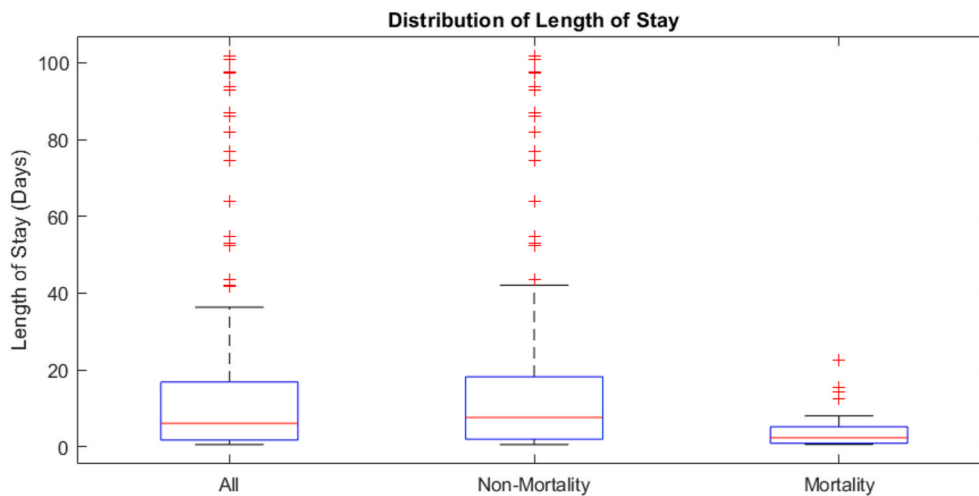


Fig. 1. Boxplots illustrating the distribution of LOS within the data.

techniques of intermittent vital sign measurement fail to capture health trends, and that continuous analysis of vital sign trends would likely improve outcomes for NICU patients. Another work [33] identified that short-term variability of heart rate (HR) and respiratory rate (RR) are strong predictors of high morbidity. As such, we chose to focus on HR and RR in this work. These two metrics are readily available in the MIMIC-III database, indicating that they are currently recorded routinely and readily in NICU environments.

To capture information about the trends in these two vital signs, we selected a 12-h period at the beginning of the relevant risk window. All HRs and RRs during this 12-h window were recorded, provided that they were greater than zero to eliminate flat-line cases and impossible negative values. Statistical analysis was then applied to quantify the variation of each vital sign during the 12-h window.

For both HR and RR, the first value, last value, minimum value, maximum value, mean value, median value, and standard deviation was calculated for inclusion in the feature vector. The first and last values were chosen as these can highlight major changes in the vital sign during the 12-h window. Minimum and maximum are used to show the most extreme values during the considered window. To represent the average vital sign, both mean and median were recorded. While mean is typically more useful, median is helpful in the case where there are significant outliers. Finally, the standard deviation is used as it is a strong indicator of variability. Where either HR or RR measurements were completely absent from a patient's record, that record was discarded and not used for training or testing.

The final feature array was as follows: *birthweight, gestational age, sex, first value for HR, last value for HR, minimum HR, maximum HR, mean HR, median HR, standard deviation of HR, first value for RR, last value for RR, minimum RR, maximum RR, mean RR, median RR, and standard deviation of RR*. The majority of these features were continuous, with the exception of sex which was binary and gestational age which was categorized into discrete bins in the database.

These features were calculated from the first 12 h for each of the considered risk windows; 3-day, 7-day, and 14-day. Cohorts varied in size for each considered risk window, due to differing levels of missingness in the data for different windows.

2.3. Balancing the dataset

Following data and feature selection, it was clear that the data was strongly unbalanced. In the cohort that met all criteria for inclusion in training and testing the 3-day NAIMS scheme, only 1.02% of the 2751 cases ended in mortality. Similarly, for 7-day and 14-day NAIMS, the mortality rates were 1.02% of 2751 cases and 1.09% of 2753 cases,

respectively. The level of imbalance can create significant overfitting issues when training a neural network, and can lead to artificially high results should the network overfit to the majority class. To prevent these issues, the non-mortality cases were undersampled by saving only the first 150 non-mortality records with the required features. Undersampling was preferred to data augmentation as it ensured that only authentic data was used in training and testing this tool for critical healthcare applications.

Following undersampling, the mortality rate in the 3-day and 7-day cohorts was 15.64% of 179 cases, while for 14-day NAIMS the mortality rate was 16.47% of 181 cases. Class weighting was used to further balance the dataset during training. Further statistical analysis of the cohorts used for training and testing each version are outlined in Tables 1 and 2 below, with 3-day, 7-day, and 14-day names hereafter denoted as NAIMS-3, NAIMS-7 and NAIMS-14 respectively.

2.4. Neural network structure

Hybrid networks have previously been used in mortality prediction for adults in one work that used extensive laboratory values and vital signs over a 48-h window, achieving reasonable AUROC of 0.834 [22]. While this shows good ability to distinguish between mortality and non-mortality cases, the dependence on long measurement windows and laboratory measurements limits the usability of the scheme for adult patients, let alone neonatal patients.

As such, in our previous work [29] we proposed a hybrid CNN-LSTM neural network, which combined multiple convolutional layers with a bidirectional LSTM layer. This structure showed high success in predicting mortality for an adult patient cohort using 51 features describing demographics, heart rate, respiratory rate, blood pressure, temperature, and blood oxygen saturation. Based on the success of the model proposed in our previous work, in this work we propose a shallower

Table 1
Characteristics of patient cohort for NAIMS-3 and NAIMS-7.

Characteristic	All patients (n = 179)	Survived (n = 151)	Died (n = 28)
Average and range of birthweights (kg)	2.13 (0.46–4.76)	2.27 (0.61–4.76)	1.39 (0.46–3.64)
Female	77 (43.02%)	71 (47.02%)	6 (21.43%)
Gestational age at birth (weeks)			
≤ 24	16 (8.94%)	5 (3.31%)	11 (39.29%)
25–28	22 (12.29%)	14 (9.27%)	8 (28.57%)
29–32	15 (8.38%)	13 (8.61%)	2 (7.14%)
33–36	87 (48.60%)	85 (56.29%)	2 (7.14%)
≥37	39 (21.79%)	34 (22.52%)	5 (17.86%)

Table 2
Characteristics of patient cohort for NAIMS-14.

Characteristic	All patients (n = 181)	Survived (n = 151)	Died (n = 30)
Average and range of birthweights (kg)	2.12 (0.46–4.76)	2.27 (0.61–4.76)	1.37 (0.46–3.64)
Female	79 (43.65%)	71 (47.02%)	8 (26.67%)
Gestational age at birth (weeks)			
≤ 24	16 (8.84%)	5 (3.31%)	11 (36.67%)
25–28	24 (13.26%)	14 (9.27%)	10 (35.72%)
29–32	15 (8.29%)	13 (8.61%)	2 (6.67%)
33–36	87 (48.07%)	85 (56.29%)	2 (6.67%)
≥37	39 (21.55%)	34 (22.52%)	5 (6.67%)

CNN-LSTM hybrid neural network, illustrated in Fig. 2, featuring a reduced hybrid structure of a single convolutional layer and single bidirectional LSTM layer. As the scheme proposed in this work has substantially fewer inputs than our previous work, the shallower and more efficient structure was sufficient to achieve strong performance.

The hybridisation of these two NN types combines the benefits of both. CNNs are well known for their ability to identify important features, while LSTM networks are well known for their ability to remember previous data in the sequence. These attributes are important in health applications, where the most important features are often not known. Deeper networks were trialled, however no improvement in performance was seen as a result of adding further layers. As such, the shallow network was used for maximum computational efficiency.

Our proposed NAIMS network uses the 17-feature vector outlined in the previous subsection as the input. This input vector is passed to the first layer, a temporal CNN layer with 128 hidden units that can be mathematically denoted as follows.

$$y_j^i = \max \left(0, \sum_{n=1}^N w_{jn}^i * x_m^{(i-1)} + b_j^i \right). \quad (1)$$

where y_j^i is the j th output feature map from the i th layer. The term w_{jn}^i ,

denotes the n th weight of the j th output feature map from the $(i - 1)$ th layer, with $n = 1, \dots, N$. The bias term b_j is the j th bias term of the i th layer. Weights and biases are updated during training using the Adam optimization algorithm. The outputs of the $(i - 1)$ th layer are denoted as $x_m^{(i-1)}$ represents the outputs of the $(i - 1)$ th layer. Finally, the convolution operation itself is denoted by the asterisk symbol (*).

The temporal CNN layer is then followed by a temporal average pooling layer, with pool and stride sizes of 2. This operation steps through the output of the CNN layer and takes the average of each pool. This results in a downsampled output, which aids in improving network efficiency and the prevention of overfitting. The addition of dropout layers was found to not improve the performance of the network, and thus only pooling was used to reduce the dimensionality of the layers.

The output from the pooling layer is then passed to a bidirectional LSTM (BiLSTM) layer with 128 hidden units. The mathematical structure of the layer is shown in Eqs. (2)–(7). As the layer is bidirectional, the data is passed through this mathematical process in both original and reversed orders. The benefit of bidirectionality is that the layer can learn from both past and future values in the sequence.

$$\tilde{c}_t = \tanh(w_c[a_{(t-1)}, x_t] + b_c) \quad (2)$$

$$f_t = \sigma(w_f[a_{(t-1)}, x_t] + b_f) \quad (3)$$

$$u_t = \sigma(w_u[a_{(t-1)}, x_t] + b_u) \quad (4)$$

$$o_t = \sigma(w_o[a_{(t-1)}, x_t] + b_o) \quad (5)$$

$$c_t = u_t \bullet \tilde{c}_t + f_t \bullet c_{(t-1)} \quad (6)$$

$$a_t = o_t \bullet \tanh(c_t) \quad (7)$$

where weights are indicated by w_c , w_f , w_u and w_o , respectively. Biases are indicated by b_c , b_f , b_u and b_o , respectively. Again, biases and weights are updated using the Adam optimization algorithm. Outputs of the previous layer are denoted as $a_{(t-1)}$, while x_t is the input to time t .

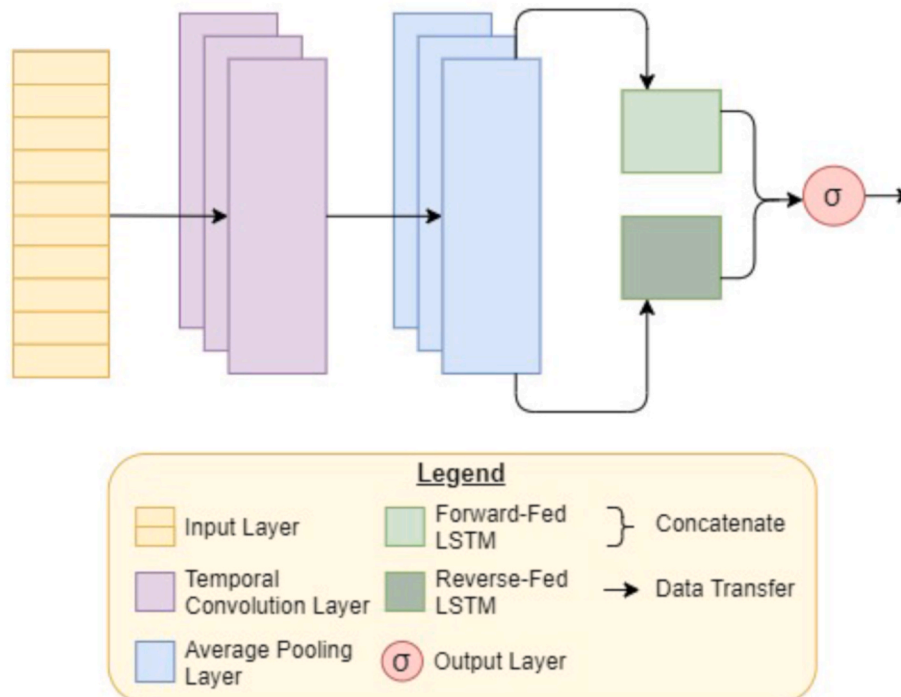


Fig. 2. Neural network structure for NAIMS.

Equations (6) and (7) are the updated cell state and layer output respectively. Element-wise multiplication is denoted by ' \odot ', while σ is the sigmoid activation function. The final layer of NAIMS is a densely-connected node utilizing sigmoid activation. Where the result of this activation is ≥ 0.5 , the patient is predicted to die within the 3-day, 7-day, or 14-day window of the respective networks. Conversely, a result < 0.5 indicates survival for that period.

2.5. Training & testing the algorithms

The NAIMS network was trained using stratified k -fold cross-validation with 5 folds, a method that splits data in 5 different ways while ensuring consistent ratios of the positive to negative cases in each split. Data was split randomly, with each record used as part of the testing set in only one of the five folds. Results obtained via cross-validation provide a more realistic view of the network performance. Due to using five folds for cross-validation, 20% of the data was used for testing in each fold, and thus remained unseen to the network while training for that fold. For training and validation, 60% and 20% of the data were used, respectively.

It is worth noting that while 10-fold cross-validation is more common in the literature, it utilizes only 10% of data for testing per fold. Given the small size of this dataset and low representation of mortality cases, this would lead to extremely few mortality cases being included in testing. Thus, 5-fold was chosen to increase the number of cases - particularly mortality cases - used in each testing run to better assess the performance of the network.

The unbalanced dataset was also accounted for when setting weights for training. Even after undersampling the non-mortality cases, mortality occurred in only 16.47% of cases. To prevent overfitting to the majority case of non-mortality, heavier weightings were placed on the importance of learning the mortality case until their relative importances were roughly equivalent. This ensured that the network would consider accurate prediction of the death and survival cases as equally important, which is essential to prevent overfitting to either case.

For each of the five folds, NAIMS was trained for 75 epochs with a single batch encompassing all training samples and with binary cross-entropy used as the loss function. This combination was found to yield the best results, with Fig. 3a-c illustrating training and validation loss over the 75 epochs for one fold of each NAIMS model. Each of these plots shows a decrease in validation loss corresponding to the decrease in testing loss, illustrating that the models are capable of generalizing to new data.

As AUROC quantifies the ability of a network to distinguish between mortality and non-mortality cases, it was considered as the primary method for evaluating network performance. As such, the training and validation AUROC was also assessed to determine the goodness of fit for

each model. As is shown in Fig. 4a-c, the validation AUROC steadily rose along with the training AUROC. The close proximity and comparable trend of the training and validation lines show that the models generalized well when applied to validation data, indicating that overfitting has not occurred.

Overall, it is clear that the models fit well to the data and are capable of generalizing well to new data. During each fold, the weights that resulted in the lowest validation loss were used for testing.

3. Results & discussions

In analysing the performance of the NAIMS networks, the key metric considered was area under the receiver-operator curve (AUROC). AUROC is the most common metric used for analysing diagnostics tools, and is calculated from the receiver-operator curve (ROC). Higher AUROC values indicate stronger ability to distinguish between the mortality and non-mortality case. Fig. 5 plots the ROC curves for NAIMS-3, NAIMS-7 and NAIMS-14, with the AUROC shown in the legend. From this graph, it is clear that all schemes achieve good AUROCS, with that NAIMS-3 unsurprisingly achieving the best result.

Area under the precision-recall curve (AUPRC) is also often considered, particularly where data is imbalanced and the predictive performance on the positive cases is highly important. AUPRC is considered with respect to the performance of a baseline random classifier, which would vary in performance depending upon the imbalance of the data. The higher the AUPRC is above the random classifier, the better its ability to distinguish between the two classes. The precision-recall curves for all NAIMS schemes are shown in Fig. 6, with AUPRC values presented in the legend. This figure clearly shows that the AUPRC for NAIMS-3 is very strong, with NAIMS-7 and NAIMS-14 also performing quite strongly. Overall, this indicates that all three models are distinguishing well between mortality and non-mortality cases.

The results obtained by NAIMS-3, NAIMS-7 and NAIMS-14 during the testing phase are further numerically summarised in Table 3. In addition to AUROC and AUPRC, several metrics were considered to evaluate the predictive accuracy. Overall accuracy (ACC), sensitivity or true positive rate (TPR), and specificity or true negative rate (TNR) were calculated for each version of NAIMS. TPR and TNR are useful metrics for an unbalanced data set, as they show the accuracy for the positive (mortality) and negative (survival) cases, respectively. This allows for analysis of the fit; if ACC, TPR and TNR are all similar values, then the network has fit equally well to both cases despite the imbalance of the data.

As shown in Table 3, NAIMS-3 performed extremely strongly. NAIMS-3 achieved an excellent AUROC of 0.9336 with strong overall accuracy and highly similar performance on both the positive and negative classes, as shown by the TPR and TNR values. Both NAIMS-7

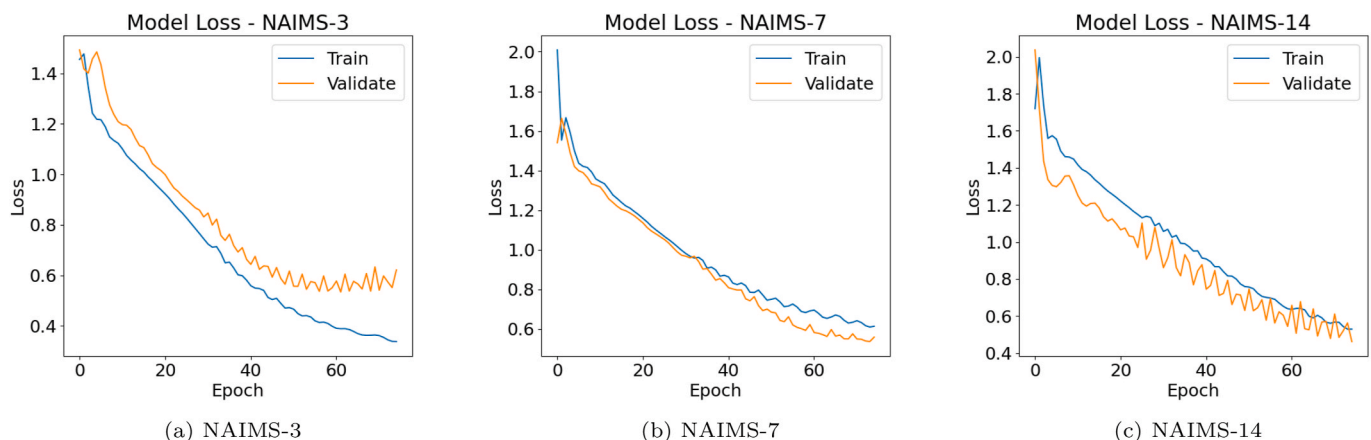


Fig. 3. Training and validation loss over epochs for NAIMS networks.

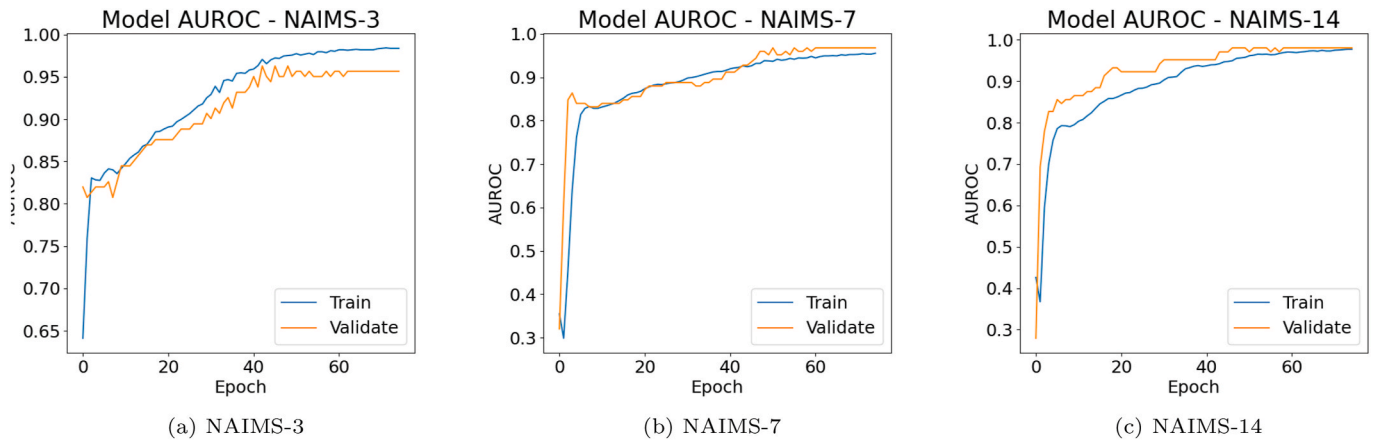


Fig. 4. Training and validation AUROC over epochs for NAIMS networks.

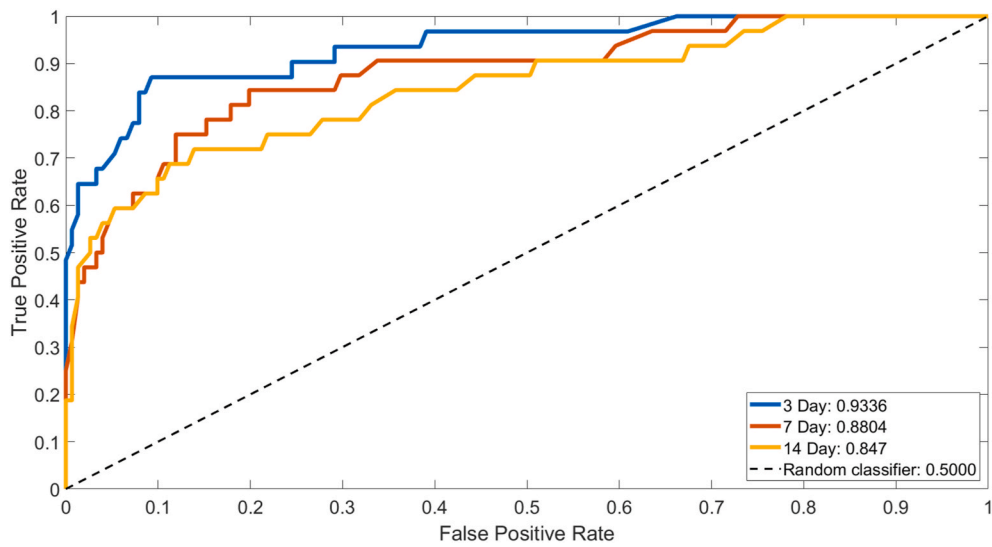


Fig. 5. ROCs for all NAIMS schemes.

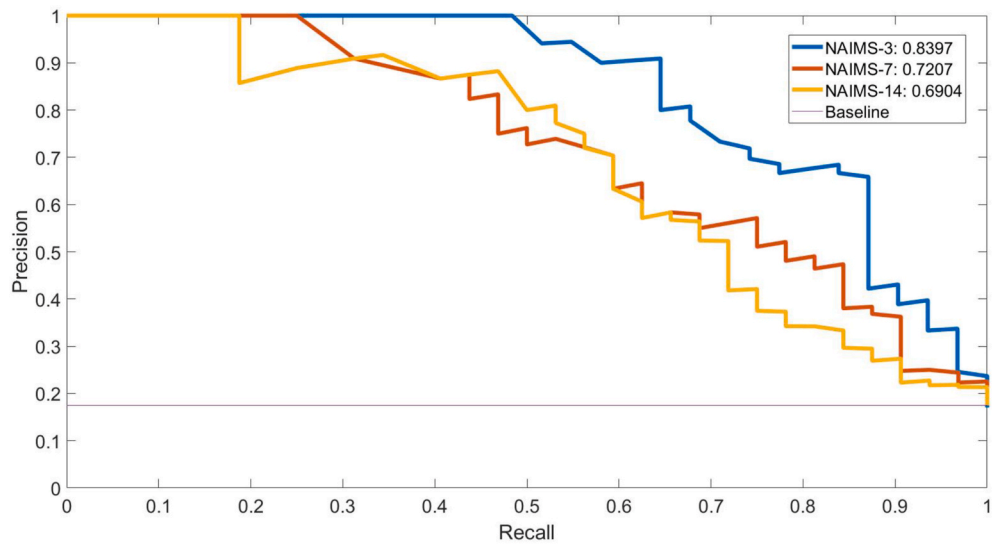


Fig. 6. Comparison of PRCs for all NAIMS schemes.

Table 3

Results obtained by NAIMS, presented as the average across the 5 folds with standard deviation in parentheses.

Scheme	ACC (%)	TPR	TNR	AUROC	AUPRC
NAIMS-3	86.81 (7.88)	0.8710 (0.0726)	0.8675 (0.0941)	0.9336 (0.0337)	0.8397 (0.0356)
NAIMS-7	80.33 (5.74)	0.8125 (0.0568)	0.8013 (0.0787)	0.8804 (0.0471)	0.7207 (0.1511)
NAIMS-14	73.77 (6.09)	0.7500 (0.0916)	0.7351 (0.0894)	0.8470 (0.0259)	0.6904 (0.0824)

and NAIMS-14 also achieved good results, with strong AUROC values and reasonable accuracy. Unsurprisingly, NAIMS-3 was the strongest performer across all metrics. This is likely due to its shorter predictive window, as the patient is more likely to be showing signs of deterioration when mortality risk is assessed for a shorter window. Meanwhile NAIMS-14 had the lowest performance, likely due to the longer window. This trend further emphasises the need for continuous short-term mortality risk assessment as an alternative or compliment to the single mortality assessment that is routinely performed at admission time in NICUs today.

It is also worth noting that the models could readily be tuned to focus more heavily on predicting the mortality case, however this would lead to reduced predictive performance for the non-mortality case and thus an increase in false alarms. This leads to a risk of alarm fatigue, a well documented phenomenon in NICU and general hospital environments [34–36] wherein healthcare workers are overwhelmed by the large number of patient health alarms and thus become desensitized to them. Alarm fatigue leads to serious risk of missing significant alarms, which has previously lead to deaths in hospitals [35].

Conversely, the models could be tuned to focus on the non-mortality case; which would incidentally increase the model accuracy as mortality rate is relatively low in NICU environments. However, identifying patients at risk of mortality is more critical than identifying patients who will survive, as the former group will require stronger intervention to maximise their chances of recovery. As such, it is preferable to obtain a lower accuracy with higher sensitivity to mortality cases and strong ability to distinguish between the two classes. Overall, NAIMS was designed to offer a balance between minimising false alarm fatigue whilst also maximising the number of true mortality risks identified.

3.1. Feature importances

An important factor for the adoption of artificial intelligence tools into clinical environments is the interpretability and explainability of the results [37–39]. One common method for interpreting the results of neural networks is SHAP (SHapley Additive exPlanations) [40], which involves the calculation of feature importances for each feature in each record when making a prediction. These can then be used to assess how a model makes predictions both on a global level and on a local level, with Fig. 7a–c illustrating the global perspective for NAIMS-3, NAIMS-7, and NAIMS-14.

In Fig. 7a–c, each data point represents the SHAP value of a particular feature as it contributes to the output of one record. The x-axis indicates the impact on the model output; that is, whether the concerned feature pushed the prediction towards non-mortality (negative) or mortality (positive). The colour scale indicates whether the value of a particular feature was low (blue) through to high (red) compared to all other instances of that feature.

From this global perspective, it is possible to interpret which features had the greatest impact on patient outcome. For example, a low median RR (denoted as RR_Med) is shown to have had a strong influence on predicting a mortality outcome across all three windows. Meanwhile, a low final value for HR (denoted as HR_Last) pushed predictions towards non-mortality. There are some results that were unexpected, such as a high STD for HR influencing a non-mortality prediction. While high variability has been previously shown to be an indicator of clinical deterioration, it may also show patient improvement if vital signs are rapidly stabilised in response to treatment. This would be an interesting area for future research.

In calculating the SHAP values, we also generated “force plots” for several local instances, which allow for inspection of how features influence individual predictions. Fig. 8a–b illustrate the force plots for two records in the database.

Fig. 8a shows the force plot for a non-mortality record which was correctly predicted by NAIMS-3, while Fig. 8b shows a mortality record which was correctly predicted. Features in blue indicate features that pushed the prediction towards a non-mortality output, while red influenced the prediction towards a mortality output. The width of the bars indicates the magnitude of the feature’s influence. As such, we can determine that the features of last HR and RR (denoted as HR_Last and RR_Last, respectively) had the largest influence towards the non-

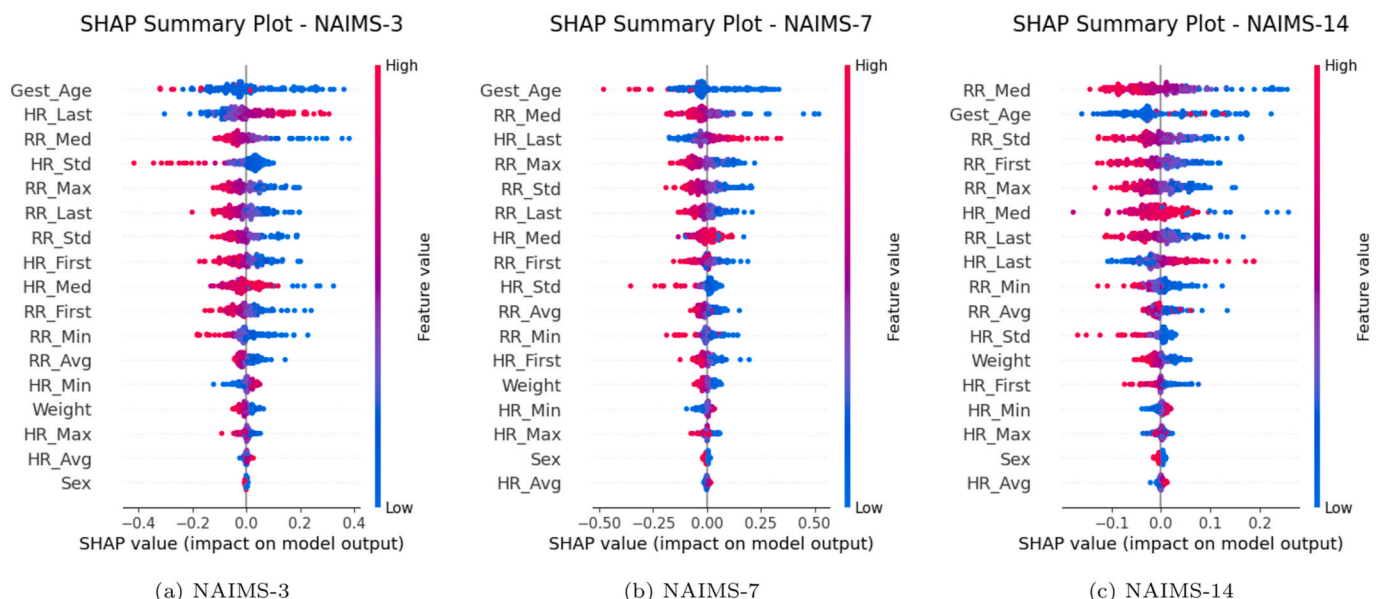


Fig. 7. SHAP summary plots for NAIMS networks.

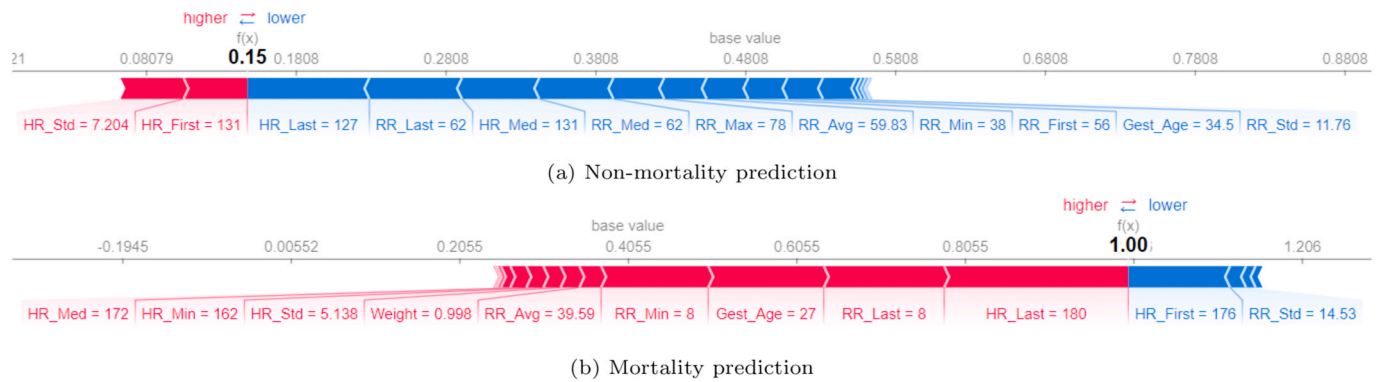


Fig. 8. Force plots for individual predictions made by the NAIMS-3 network.

mortality prediction in Fig. 8a and the mortality prediction in Fig. 8a. Local models such as these SHAP force plots could readily be used by clinicians to better interpret the outputs of the NAIMS network, which would aid in building trust in the model.

Overall, SHAP analysis serves to illustrate the global and local performance of the NAIMS model. The figures presented in this subsection clearly illustrate the influence that different features had on decision-making across the entire dataset, as well as highlighting how decisions were made in several local samples.

3.2. Comparison to previous works

In this section, we compare the results achieved by the NAIMS schemes to works presented in the literature. Table 4 compares the AUROCs of recent works, and includes descriptions of the features and measurement windows considered in each of the included works. The majority of previous works did not include ACC, TPR, TNR and/or AUPRC values in their analyses, so these have been excluded from the table.

The results presented in Table 4 indicate that NAIMS-3 outperforms all previous works in the literature, achieving a significantly higher AUROC than all previous works. This high-performing network also has several other advantages over existing schemes, including the ability to perform mortality risk assessment based on any 12-h window of data during the patient’s stay.

NAIMS-7 performs comparably to previous works, outperforming multiple schemes. NAIMS-14 performs comparably to the work presented by Houweling et al. [15], however does not perform as strongly as much of the literature. It is likely that access to additional training

data would improve the performance of all NAIMS networks, and indeed this would be the next step required to work towards implementation of these schemes in real healthcare environments.

While the proposed NAIMS networks depend upon more features than some previous works, 14 of the 17 features are easily derived from temporal HR and RR data. Conversely, several works [15,17] depend on variables that are completely distinct from each other and thus require more extensive acquisition and calculation. Furthermore, the work presented by Houweling et al. [15] depends upon subjective metrics assessing the baby’s appearance, rather than on tangible measurements. In the work presented by Jaskari et al. [18] only 14 variables are directly mentioned, however the dependence on SNAP-II and SNAPPE-II introduce many additional dependencies. Of the previous works included in Table 4, only the scheme presented by Jaskari et al. [18] could be updated on an ongoing basis during the patient stay. All other works depend on variables that are static and are measured immediately post-birth.

Another significant advantage of all NAIMS schemes is the ability to be updated regularly and automatically, which allows for easier identification of trends in the patient’s health. Sex and birthweight are determined at birth, and gestational age could be automatically updated as the baby ages. HR and RR statistical values can be automatically calculated from monitoring equipment, or from manual data entries made by healthcare staff. Meanwhile, the only other scheme in the literature that is designed in such a way that mortality risk could be updated during the stay [18] still depends upon variables that would realistically make this challenging. Namely, it depends upon SNAP-II and SNAPPE-II scores which introduce a direct dependency on parameters such as PO₂/FiO₂, base excess, and urine output. Such parameters

Table 4
Performance of NAIMS-3, NAIMS-7, NAIMS-14 and other schemes from the literature.

Scheme	Cohort Size	No. Features	Description of Features	Data Acquisition Window	Algorithm Type	AUROC
Cooper [19]	10,051	284	Birth characteristics, laboratory test results, treatments received, existing conditions	Varied - all available data from the patient stay used	Superlearner (14 ML algorithms)	0.91
Podda [17] (Best Model)	29,557	12	Birth characteristics, demographics, existing conditions, treatments received, maternal characteristics, maternal treatments received	Varied - used maternal health pre-birth, plus measurements from the first 5 min post-birth	Densely-Connected Neural Network	0.9136
Houweling [15] (Post-Birth Model)	51,374	10	Birth characteristics, maternal characteristics, condition of the baby by visual inspection	Varied - some information obtained pre-birth and during birth, plus 5 min post-birth	Logistic regression	0.85
Medvedev [16] (UK Cohort)	110,176	3	Birth weight, admission oxygen saturation, highest respiratory support within 24 h	24 (from admission)	Logistic regression	0.8903
Medvedev [16] (Gambia Cohort)	550	3	Birth weight, admission oxygen saturation, highest respiratory support within 24 h	24 h	Logistic regression	0.8082
Jaskari [18] (Best Model)	977	14	Vital signs, demographics, SNAP-II and SNAPPE-II scores	36 h	Random Forest	0.922
NAIMS-3	179	17	Gestational age, birthweight, sex, vital signs	12 h (any window)	CNN-LSTM	0.9336
NAIMS-7	179	17	Gestational age, birthweight, sex, vital signs	12 h (any window)	CNN-LSTM	0.8804
NAIMS-14	181	17	Gestational age, birthweight, sex, vital signs	12 h (any window)	CNN-LSTM	0.8470

are substantially more complex to measure than vital signs, and thus would introduce a higher burden on healthcare workers.

Compared to previous works using hybrid neural networks to predict mortality risk in other populations, NAIMS performs strongly. Two works have previously considered CNN-LSTM networks for mortality prediction in adults. In Ref. [22], a deep CNN-LSTM is used to predict mortality from 37 laboratory values and vital signs, achieving an AUROC of 0.836. Meanwhile, our own previous work employed a CNN-LSTM with multiple CNN layers to predict mortality from 51 features describing demographics and vital signs, with the strongest model achieving an AUROC Of 0.884. Comparatively, NAIMS is the shallowest CNN-LSTM model, has the fewest input features, and achieves an AUROC of up to 0.9336. The results of this work clearly demonstrate that a well-designed hybrid neural network with carefully selected features is suitable for a neonatal patient cohort, while previous works validate their use on adult patient cohorts.

3.3. Limitations

This work serves as a pilot study on the use of CNN-LSTM networks to predict mortality risk from basic demographics and vital signs. There are several limitations that must be acknowledged. The primary limitation is the small quantity of mortality data. While MIMIC-III is perhaps the largest open-access database for clinical research, very few NICU stays resulted in mortality. Prior works have had access to larger databases, although it is worth noting that the parameters included in those databases were vastly different than those included in this work. While this pilot study shows highly promising results, further validation on a larger database would be required to conclusively prove the performance of our proposed NAIMS scheme.

The high imbalance in the available dataset also resulted in a need to perform undersampling and class weighting to enhance the ability of the network to learn from the data. This reduced the size of the non-mortality dataset. While necessary in this context, it would be preferable to obtain larger quantities of data in both classes so that data balancing could be performed whilst maintaining a large database.

In our future work, we aim to address these limitations by pursuing partnerships with local hospitals to clinically validate the NAIMS tool on large and diverse patient populations.

4. Conclusion

In this work, we have proposed the NAIMS shallow hybrid neural network structure that utilizes temporal convolution, pooling, and long short-term memory layers. NAIMS was then trained and tested for predicting mortality risk within the following 3, 7, and 14 day periods, resulting in NAIMS-3, NAIMS-7, and NAIMS-14, respectively.

It was shown that our NAIMS schemes perform well when compared to existing schemes in the literature. In particular, NAIMS-3 outperforms all works in the literature, highlighting the benefit of shorter-term mortality risk assessment. All NAIMS scores can be recalculated regularly and automatically, allowing for ongoing analysis of the patient's condition during the NICU stay.

Furthermore, NAIMS uses a short, 12-h window of temporal data to make its predictions, allowing the first prediction to be made within half a day of admission, without needing knowledge of maternal condition prior to birth. NAIMS also depends only upon simple features that are readily available in the NICU environment already. This simplicity enables regular and automatic recalculation of mortality risk during the stay, which in turn enables healthcare workers to monitor a patient's health trends and response to any treatments.

As a result of these benefits and the strong ability to distinguish between mortality and non-mortality cases, we suggest that our NAIMS schemes are suitable for use in predicting mortality risk in NICU environments. In our own future work, we aim to take these algorithms to clinical trials to further validate and improve the NAIMS schemes in

terms of accuracy and usability. Additionally, we aim to conduct research into the prediction of other outcomes in NICU environments, such as the onset of sepsis and patient length of stay.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Australian Government Research Training Program Scholarship.

References

- [1] L. Liu, S. Oza, D. Hogan, et al., Global, regional, and national causes of under-5 mortality in 2000-2013: an updated systematic analysis with implications for the Sustainable Development Goals, *Lancet* 388 (10063) (2016) 3027–3035, [https://doi.org/10.1016/S0140-6736\(16\)31593-8](https://doi.org/10.1016/S0140-6736(16)31593-8).
- [2] J.E. Lawn, M.G. Gravett, T.M. Nunes, et al., Global report on preterm birth and stillbirth (1 of 7): definitions, description of the burden and opportunities to improve data, *BMC Pregnancy Childbirth* 10 (Suppl 1) (2010), <https://doi.org/10.1186/1471-2393-10-S1-S1>.
- [3] March of Dimes, Pmnc, Save the children. *Born Too Soon: the Global Action Report on Preterm Birth*, Tech. Rep., World Health Organization, Geneva, Switzerland, 2012.
- [4] W. Harrison, D. Goodman, Epidemiologic trends in neonatal intensive care, 2007-2012, *JAMA, Pediatrics* 169 (9) (2015) 855–862, <https://doi.org/10.1001/jamapediatrics.2015.1305>.
- [5] J.H. Park, Y.S. Chang, S.Y. Ahn, et al., Predicting mortality in extremely low birth weight infants: Comparison between gestational age, birth weight, Apgar score, CRIB II score, initial and lowest serum albumin levels, *PLoS One* 13 (2) (2018), e0192232, <https://doi.org/10.1371/journal.pone.0192232>.
- [6] G. Parry, J. Tucker, W. Tarnow-Mordi, Crib II: an update of the clinical risk index for babies score, *Lancet* 361 (9371) (2003) 1789–1791, [https://doi.org/10.1016/S0140-6736\(03\)13397-1](https://doi.org/10.1016/S0140-6736(03)13397-1).
- [7] W. Tarnow-Mordi, S. Ogston, A.R. Wilkinson, et al., Predicting death from initial disease severity in very low birthweight infants: a method for comparing the performance of neonatal units, *Br. Med. J.* 300 (6740) (1990) 1611–1614, <https://doi.org/10.1136/bmj.300.6740.1611>.
- [8] D.K. Richardson, J.E. Gray, M.C. McCormick, et al., Score for Neonatal Acute Physiology: a physiologic severity index for neonatal intensive care, *Pediatrics* 91 (3) (1993) 617–623.
- [9] D.K. Richardson, C.S. Phibbs, J.E. Gray, et al., Birth weight and illness severity: independent predictors of neonatal mortality, *Pediatrics* 91 (5) (1993) 969–975.
- [10] D.K. Richardson, J.D. Corcoran, G.J. Escobar, et al., SNAP-II and SNAPPE-II: simplified newborn illness severity and mortality risk scores, *J. Pediatr.* 138 (1) (2001) 92–100, <https://doi.org/10.1067/mpd.2001.109608>.
- [11] R.F. Maier, M. Rey, B.C. Metzke, et al., Comparison of mortality risk: a score for very low birthweight infants, *Arch. Dis. Child. Fetal Neonatal Ed.* 76 (3) (1997) F146, <https://doi.org/10.1136/fn.76.3.F146>. LP – F151.
- [12] H. García, R. Villegas-Silva, D. Villanueva-García, et al., Validation of a prognostic index in the critically ill newborn, *Revista de investigación clínica; organo del Hospital de Enfermedades de la Nutrición* 52 (4) (2000) 406–414.
- [13] B. Garg, D. Sharma, N. Farahbakhsh, Assessment of sickness severity of illness in neonates: review of various neonatal illness scoring systems, *J. Matern. Fetal Neonatal Med.* 31 (10) (2018) 1373–1380, <https://doi.org/10.1080/14767058.2017.1315665>.
- [14] S.S. Harsha, B.R. Archana, SNAPPE-II (score for neonatal Acute Physiology with perinatal extension-II) in predicting mortality and morbidity in NICU, *J. Clin. Diagn. Res.* 9 (10) (2015) SC10–SC12, <https://doi.org/10.7860/JCDR/2015/14848.6677>.
- [15] T.A.J. Houweling, D. van Klaveren, S. Das, et al., A prediction model for neonatal mortality in low- and middle-income countries: an analysis of data from population surveillance sites in India, Nepal and Bangladesh, *Int. J. Epidemiol.* 48 (1) (2018) 186–198, <https://doi.org/10.1093/ije/dyy194>.
- [16] M.M. Medvedev, H. Brotherton, A. Gai, et al., Development and validation of a simplified score to predict neonatal mortality risk among neonates weighing 2000 g or less (NMR-2000): an analysis using data from the UK and the Gambia, *The Lancet Child & Adolescent Health* 4 (4) (2020) 299–311, [https://doi.org/10.1016/S2352-4642\(20\)30021-3](https://doi.org/10.1016/S2352-4642(20)30021-3).
- [17] M. Podda, D. Bacciu, A. Micheli, et al., A machine learning approach to estimating preterm infants survival: development of the Preterm Infants Survival Assessment (PISA) predictor, *Sci. Rep.* 8 (1) (2018) 13743, <https://doi.org/10.1038/s41598-018-31920-6>.
- [18] J. Jaskari, J. Myllärinen, M. Leskinen, et al., Machine Learning Methods for Neonatal Mortality and Morbidity Classification, *IEEE Access*, 2020, <https://doi.org/10.1109/ACCESS.2020.3006710>.

- [19] J.N. Cooper, P.C. Minneci, K.J. Deans, Postoperative neonatal mortality prediction using superlearning, *J. Surg. Res.* 221 (2018) 311–319, <https://doi.org/10.1016/j.jss.2017.09.002>.
- [20] R.O. Deliberato, G.G. Escudero, L. Bulgarelli, et al., SEVERITAS: an externally validated mortality prediction for critically ill patients in low and middle-income countries, *Int. J. Med. Inf.* 131 (2019) 103959, <https://doi.org/10.1016/j.ijmedinf.2019.103959>.
- [21] R. Yu, Y. Zheng, R. Zhang, et al., Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients, *IEEE Journal of Biomedical and Health Informatics* 24 (2) (2020) 486–492, <https://doi.org/10.1109/JBHI.2019.2916667>.
- [22] T. Alves, A. Laender, A. Veloso, et al., Dynamic prediction of ICU mortality risk using domain adaptation, in: Proc. Of the 2018 IEEE International Conference on Big Data, 2019, pp. 1328–1336doi, <https://doi.org/10.1109/BigData.2018.8621927>.
- [23] M.A. Zahid, J. Lee, Mortality prediction with self normalizing neural networks in intensive care unit patients, Proc. of the 2018 IEEE EMBS International Conference on Biomedical and Health Informatics (2018) 226–229doi, <https://doi.org/10.1109/BHI.2018.8333410>.
- [24] A.E. Johnson, R.G. Mark, Real-time mortality prediction in the intensive care unit, *AMIA Annual Symposium Proceedings 2017* (2017) 994–1003.
- [25] R.J. Delahanty, D. Kaufman, S.S. Jones, Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients, *Crit. Care Med.* 46 (6) (2018) e481–e488, <https://doi.org/10.1097/CCM.0000000000003011>.
- [26] F. Miao, Z.D. Liu, J.K. Liu, et al., Multi-sensor fusion approach for cuff-less blood pressure measurement, *IEEE Journal of Biomedical and Health Informatics* 24 (1) (2020) 79–91, <https://doi.org/10.1109/JBHI.2019.2901724>.
- [27] K. Yu, M. Zhang, T. Cui, et al., Monitoring ICU mortality risk with A long short-term memory recurrent neural network, *Pacific Symposium on Biocomputing 25* (2020) 103–114, https://doi.org/10.1142/9789811215636_0010.
- [28] J. Fagerström, M. Bång, D. Wilhelms, et al., LiSep LSTM: a machine learning algorithm for early detection of septic shock, *Sci. Rep.* 9 (1) (2019) 15132, <https://doi.org/10.1038/s41598-019-51219-4>.
- [29] S. Baker, W. Xiang, I. Atkinson, Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach, *Sci. Rep.* 10 (1) (2020) 21282, <https://doi.org/10.1038/s41598-020-78184-7>.
- [30] A.E. Johnson, T.J. Pollard, L. Shen, et al., MIMIC-III, a freely accessible critical care database, *Scientific Data* 3 (2016) 160035, <https://doi.org/10.1038/sdata.2016.35>.
- [31] E. Nagy, H. Orvos, G. Bárdos, et al., Gender-related heart rate differences in human neonates, *Pediatr. Res.* 47 (6) (2000) 778–780, <https://doi.org/10.1203/00006450-200006000-00016>.
- [32] N. Kumar, G. Akangire, B. Sullivan, et al., Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront, *Pediatr. Res.* 87 (2) (2020) 210–220, <https://doi.org/10.1038/s41390-019-0527-0>.
- [33] S. Saria, A.K. Rajani, J. Gould, et al., Integration of early physiological responses predicts later illness severity in preterm infants, *Sci. Transl. Med.* 2 (48) (2010), <https://doi.org/10.1126/scitranslmed.3001304>, 65–48.
- [34] T. Tanner, The problem of alarm fatigue, *Nursing for Women's Health* 17 (2) (2013) 153–157, <https://doi.org/10.1111/1751-486X.12025>.
- [35] S. Sendelbach, M. Funk, Alarm fatigue: a patient safety concern, *AACN Adv. Crit. Care* 24 (4) (2013) 378–386, <https://doi.org/10.4037/NCI.0b013e3182a903f9>.
- [36] C.F. Poets, Reducing alarms in the NICU, *Arch. Dis. Child. Fetal Neonatal Ed.* 103 (4) (2018) F297, <https://doi.org/10.1136/archdischild-2017-314259>. LP – F298.
- [37] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, *Neural Comput. Appl.* (2019) 1–15.
- [38] M. Reyes, R. Meier, S. Pereira, et al., On the interpretability of artificial intelligence in radiology: challenges and opportunities, *Radiol. Artif. Intell.* 2 (3) (2020), e190043, <https://doi.org/10.1148/ryai.2020190043>.
- [39] L.H. Gilpin, D. Bau, B.Z. Yuan, et al., Explaining explanations: an overview of interpretability of machine learning, in: 2018 IEEE 5th Int. Conf. Data Sci. Adv. Anal., 2018, pp. 80–89, <https://doi.org/10.1109/DSAA.2018.00018>.
- [40] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, 2017 [arXiv:1705.07874](https://arxiv.org/abs/1705.07874).