Machine learning for mortality risk prediction with changing patient demographics

WAINWRIGHT, Richard and SHENFIELD, Alex <http://orcid.org/0000-0002-2931-8077>

Machine learning for mortality risk prediction with changing patient demographics

Richard Wainwright
Department of Engineering and Mathematics
Sheffield Hallam University
Sheffield, UK
ric.wainwright@theinsightsfamily.com

Alex Shenfield
Department of Engineering and Mathematics
Sheffield Hallam University
Sheffield, UK
a.shenfield@shu.ac.uk

***Abstract*—Over the last 25-30 years there has been significant work carried out in producing risk prediction models for patients admitted to intensive care units. The most recent of these models in widespread use is the Intensive Care National Audit and Research Centre (ICNARC) model developed in 2007 which uses data from more than 230,000 admissions to UK intensive care units to develop and validate a UK based model outperforming other approaches. However, as with the majority of risk prediction models, the ICNARC model struggles with changing patient cohort demographics (such as the aging populations seen currently in the western world) and requires periodic recalibration.**

This paper introduces a machine learning pipeline for developing mortality prediction models and uses it to train a variety of ML models. The top performing of these outperform current commonly used mortality risk prediction models such as APACHE-II, SAPS-II, and the ICNARC model. This machine learning pipeline is then extended to allow continuous retraining via online learning. The results show that it is possible to retrain our model at different intervals to deal with varying patient demographics - improving model performance across a range of different patient cohort scenarios.

***Index Terms*—Machine Learning, Online Learning, Risk Prediction, Mortality Prediction, ICU**

I. Introduction

Over the last decade, UK National Health Service (NHS) waiting times have increasingly failed to meet targets. This is now compounded by challenges in recruitment and retention of nursing and other hospital staff meaning that intensive care unit (ICU) staffing ratios are often well below recommended levels [1]. To mitigate the impact of this staffing crisis, prediction of mortality risk in the ICU environment is of increasing importance to allow stratification of patients and targeted support.

There has been significant work carried out in developing risk prediction models for patients admitted to intensive care units [2], [3]. In the UK, the most recent of these is the Intensive Care National Audit and Research Centre (ICNARC) model which uses data from 163 intensive care units across the UK (covering 231,900 admissions) to develop and validate a model outperforming previous approaches such as the SAPS-II (Simplified Acute Physiology Score) [4] and APACHE-II (Acute Physiology and Chronic Evaluation) [5] scores.

However, a key disadvantage of these mortality risk prediction models is that they are developed using retrospective patient cohorts, with patient demographics and the effectiveness of medical treatments changing over time. This means that periodically these risk prediction models require recalibration to take into account changes in treatment and changes to patient cohorts.

In this paper the application of several different machine learning (ML) techniques for accurately predicting mortality risk in the ICU is discussed and their effectiveness compared. The most effective of these is then taken forward to be used in an online retraining framework that is capable of dealing with changing patient demographics. The key contributions of this research are:

1) Evaluation of existing methods used for mortality risk prediction in ICU.
2) Development of an effective ML pipeline to provide an accurate prediction of mortality risk to enhance decision support by medical practitioners.
3) Investigation of the effects of online retraining of ML models to allow for the adaption of risk models to different patient cohorts.

The remainder of the paper is structured as follows. Section II evaluates the current approaches and the existing state-of-the-art research, section III describes the dataset used in this study and the methodology, and section IV describes the results of the experiments. Section V then presents our conclusions and ideas for further work.

II. Related work

Even though there are multiple new methods for determining mortality within ICU, the APACHE II and SAPS II scores continue to be the most used point-based schemes worldwide [6]. Similarly, the Sequential Organ Failure Assessment (SOFA) score is used in some parts of the world as a mortality risk assessment tool even though it was developed to assess sepsis risk [7]. Some common limitations that are associated with these tools that have been detailed in the literature are:

1) There has been a decrease in performance over time. [8] indicated that SAPS II was not within calibration tolerance by 2005.
2) There have been calibration issues with both the APACHE II and SOFA scores when applying them to new patient cohorts [9].

979-8-3503-1017-7/23/$31.00

3) [9] and [10] noted that several of these mortality risk assessment tools were not very reliable for patients within Europe or Singapore as they were not developed with data from these patient cohorts.
4) Some variables which are required to provide a score are difficult to obtain, especially when patients are admitted into critical care situations. For many cases, the data might not be available because it requires expensive laboratory pathology tests and full patient medical history.

These limitations have led to researchers exploring alternative approaches for mortality prediction. The resurgence of ML techniques has provided some promising preliminary results in this problem domain. Furthermore, ML models are comparatively easy to update, retrain, and re-calibrate for different patient cohorts and as patient cohorts evolve over time [10]. Traditional approaches to mortality prediction often only capture a single time period; this approach misses out on valuable insights and data that could improve the models accuracy, precision, or recall as treatments and patient demographics change over time. Including online learning in developed ML models means that predictive tools can learn from new examples in real-time, ensuring that the model constantly generalises well to the populations it is applied to, even as environmental factors, operations, and medicines change.

For practical application in an ICU setting, a mortality risk prediction model should only use vital signs that can be continually monitored and should allow the doctor to see how the risk changes. [11] have developed a model using only vital signs; however, their results demonstrated limited predictive capability in determining mortality outcomes, with an area under the receiver operating curve (AUROC) of 0.65. They also used a combination of vital signs and additional features culminating in a higher AUROC of 0.85 when the data was combined with the SAPS II score and patient demographic information.

Throughout the literature, there are several ML techniques used to consider the prediction of mortality. For example, [12] used a combination of convolutional layers and recurrent layers to predict mortality on a subset of the MIMIC-II dataset [13]. This data was gathered over the first 48-hours of patients admitted to ICU. Their combined CNN-LSTM achieved an overall AUROC of 0.836 and outperformed the use of convolutional neural networks (CNNs) or long short term memory (LSTM) based recurrent neural networks alone. [14] used the MIMIC-III dataset [15] and trained multiple bi-directional LSTM models on different time windows to determine mortality risk, taking potential complications into account. Using a forty-eight hour time window, they obtained an AUROC score of 0.885 using a bi-directional LSTM.

However, there has been relatively little focus on predicting mortality at admission (within the first 24 hours) to the ICU. Although this approach doesn't take into account potential complications that occur after admission, it provides valuable information during the triage process to allow early identification of patients that may require additional monitoring. One example of this is [16], who used an artificial neural network (ANN) and the JADE optimisation algorithm [17] to obtain an accuracy of over 90% when at decision criteria between 30–80%, with an AUROC score of 0.932 on a dataset of single-site ICU admissions from the UK. In the US, the AIMS scheme [18] uses a hybrid CNN-LSTM network trained on a combination of age, gender, and a selection of statistical parameters obtained within the first 24 hours of admission into the ICU and extracted from the MIMIC-III dataset [15]. In the AIMS system, risk predictions are generated for the next 3-day, 7-day, and 14-day windows. AIMS achieved an AUROC score of 0.884–0.858 depending on the predictive window.

III. Methodology and experimental design

A. ICNARC Dataset

The research in this paper was undertaken using the ICNARC dataset that was collected at the North Middlesex University Hospital cluster between January 1st 2012 and April 30th 2014. The dataset consists of 13,494 patient records, where each row corresponds to a patient admitted into the ICU. There is no missing data in the dataset.

The dataset is comprised of 28 physiological features (shown in Table I which are obtained in the first 24 hours of admission to the ICU. As well as the physiological features, there is some additional patient information collected – including patient age at the time of admission into the ICU, whether the patient had CPR within 24 hours of admission, the location of the patient before the admission (which is often referred to as the source), and whether the patient was intubated during the first 24 hours.

TABLE I: Features of the ICNARC dataset

	Used Variable
1	Anonymised Unit Identifier
2	Age in years at last birthday
3	Gender
4	Residence Prior to admission
5	Prior Dependency
6	Severe Liver Disease
7	Haematological Malignancy
8	Metastatic Disease
9	Severe Respiratory Disease and Home Ventilation
10	Immunocompromise
11	Cardiovascular Disease
12	Renal disease
13	CPR within 24 hours prior
14	Primary reason for admission
15	ICNARC Diagnostic Category
16	Condition Description
17	Type of Admission
18	Mechanically Ventilated at admission
19	Highest level of care received in unit within 24 hours
20	Basic respiratory support
21	Advanced respiratory support
22	Basic cardiovascular support
23	Advanced cardiovascular support
24	Renal support whilst in unit
25	Neurological support whilst in unit
26	Gastrointestinal support whilst in unit
27	Dermatological support whilst in unit
28	Liver support whilst in unit

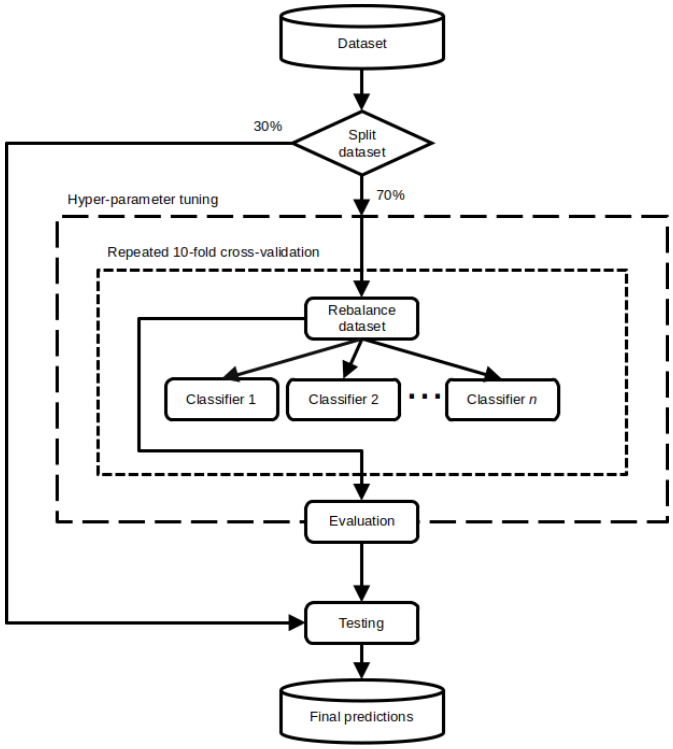


Fig. 1: Machine learning pipeline for development of a mortality risk prediction tool.

B. Developing a strong baseline

We develop a strong ML mortality risk prediction tool baseline using the ML pipeline framework shown in Figure 1. The framework is made up of the following steps:

1) The complete dataset is split into 70% training and 30% held-out blind-fold validation sets with stratification on outcome.
2) The training data is then used in hyperparameter tuning for several benchmark classification models using a repeated K-fold stratified cross-validation approach.
3) Each fold of the training data is rebalanced whilst training using SMOTE [19] to overcome the acute class imbalance (i.e. out of 13,494 patients, 11,825 survived and 1,668 died), with the validation set in each fold left alone.
4) The final models are then tested using the 30% unseen data to generate final performance scores.

In this paper, K-Nearest Neighbour (KNN), Linear Support Vector Machines (SVMs), Radial Basis Function SVMs, Gaussian Process Models, Decision Trees, Random Forests, ANNs, AdaBoost, Naïve Bayes, and QDA are investigated for use as a baseline. Each classifier is trained and evaluated using repeated stratified 10-fold cross-validation (over 10 runs of the cross-validation process), with averages of the performance metrics calculated for each classifier. Stratification techniques are used to ensure that the training and validation sets reflect the overall class imbalance of the data.

The machine learning models are evaluated using accuracy, precision, recall, and F1-score (see Table II) [20], as well as the area under the receiver operating characteristic curve (AUROC).

TABLE II: Performance metrics

Performance Metric	Definition
Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$
F1-Score	$\frac{(2\times Precision\times Recall)}{(Precision+Recall)}$

C. Online training of ML models

As discussed in section II, traditional approaches to mortality prediction models often start to perform poorly as patient cohorts and treatments change. A more effective solution is to adapt the predictive model at a local level to deal with evolving population demographics and available medical treatments. To this point, an efficient solution would be that each hospital has their own trained model that can be maintained and updated in real time.

Online learning (also known as real-time ML) is the process of continuously training a model in real-time as new data becomes available. This new data is used to update the trained parameters of the ML algorithm to find the best result for a predefined performance metric.

Event driven architectures are common ways of deploying real-time capable ML models in a production setting. The continuous flow of data through a data stream is given to the model, and the model training pipeline will handle data transformations and enrichments to ensure that the data is consistent and ready to be utilised to retrain the model.

This paper will evaluate the application of online learning to changing patient demographics by utilising the classification pipeline proposed in section III-B. This work uses subsets of the ICNARC dataset, and full details of the experimental setup and the results obtained can be found in section IV-B.

IV. Results

A. ICNARC mortality prediction

The ICNARC dataset (outlined in section III-A) was used to train a variety of ML algorithms using the ML pipeline proposed in section III-B. Performance of these classifiers was evaluated using multiple commonly used performance metrics to determine the suitability of the classifiers to the mortality prediction task. The results of this training are presented in Table III, with scores for precision, recall, F1-score, accuracy, and the AUROC presented to allow for easy comparison against results presented in the literature. These scores are the mean and standard deviation of 10 independent runs of 10-fold cross-validation and the best results obtained are highlighted in bold.

Comparing the results from Table III to those described in section II, it can be seen that the Random Forest, Linear SVM,

TABLE III: Classifier comparison on the ICNARC dataset

Classifier	Precision	Recall	F1-Score	Accuracy	AUROC
K-Nearest Neighbour	0.70 (0.02)	0.25 (0.02)	0.72 (0.04)	0.37 (0.02)	0.78 (0.02)
Linear SVM	0.82 (0.01)	**0.40 (0.02)**	0.83 (0.03)	**0.54 (0.02)**	0.91 (0.01)
RBF SVM	0.60 (0.02)	0.21 (0.01)	0.82 (0.04)	0.33 (0.02)	0.76 (0.02)
Gaussian Process	0.82 (0.01)	**0.40 (0.01)**	0.83 (0.03)	**0.54 (0.01)**	0.89 (0.01)
Decision Tree	0.77 (0.01)	0.32 (0.02)	0.77 (0.04)	0.45 (0.02)	0.77 (0.02)
Random Forest	0.82 (0.01)	**0.40 (0.02)**	**0.87 (0.03)**	**0.54 (0.02)**	**0.92 (0.01)**
Neural Network	0.80 (0.08)	0.39 (0.09)	0.78 (0.13)	0.50 (0.06)	0.89 (0.02)
AdaBoost	**0.83 (0.01)**	**0.40 (0.02)**	0.83 (0.03)	**0.54 (0.02)**	0.91 (0.01)
Naive Bayes	0.81 (0.01)	0.36 (0.03)	0.70 (0.04)	0.48 (0.03)	0.85 (0.02)
QDA	0.81 (0.02)	0.37 (0.03)	0.72 (0.04)	0.49 (0.03)	0.87 (0.01)

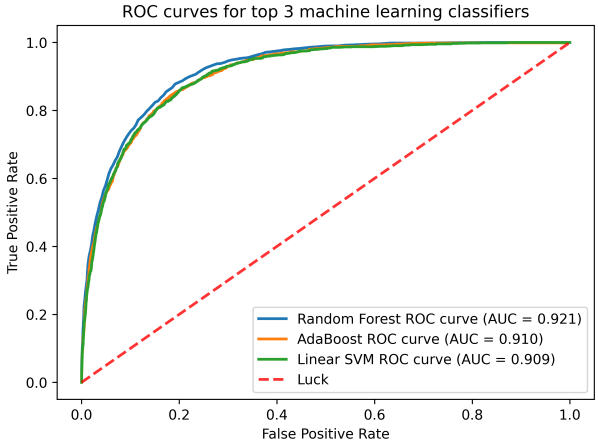
and AdaBoost classifiers obtain comparable overall AUROC scores (0.92, 0.91, and 0.91, respectively) to that presented in [16] (0.93), and significantly better performance than either the benchmark ICNARC score (0.80) or the APACHE-II score (0.83) on this dataset. The overall accuracy scores were slightly lower than those reported in [16]; however, there was greater imbalance in the data set used in this study as well as the dataset covering multiple ICU sites.

Figures 2a and 2b show the ROC curves for the top 3 classifiers and the APACHE-II score, respectively. As you can see, all 3 of the top performing classifiers demonstrate excellent predictive capability across a range of operating thresholds. Figure 3 also shows the confusion plots for the top 3 classifiers identified above. Data in these confusion plots is taken from the unseen data (i.e., the validation partitions) of a single representative run of the repeated 10-fold cross-validation process.

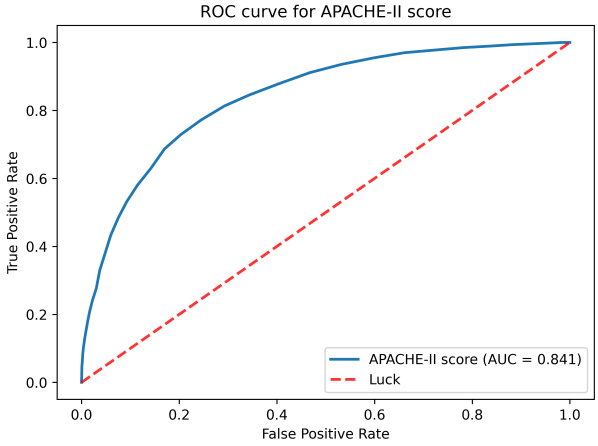
B. Dealing with changing patient demographics via online learning

This section demonstrates that online retraining of ML models, in real-time (i.e. as new data becomes available), can be used as an effective way to develop mortality risk prediction tools that can deal with changing patient cohorts. Two patient cohort scenarios are investigated:

1) Patient cohort by date of admission to ICU: In this scenario the ICNARC data discussed in section III-A is partitioned based on the month of admission into the ICU to investigate the impact of seasonality on classifier performance and to show how retraining the model as new data becomes available helps to ensure that the classifier remains useful. Figure 4 shows how classifier performance degrades due to seasonality. In this figure, a random forest classifier is trained on ICU admission data from a given month and tested on both a held out validation set from that month and the data from other months (using repeated 10-fold cross-validation). We can see that, as the testing months get further away from the month that the classifier was trained on, the accuracy and AUROC measure reduce until they get to a minimum value around 5 and 6 months away from the training data (this corresponds to the greatest difference in seasonal conditions). A significant drop in mean accuracy can be observed at 5 months away from the training data (though this is not as obvious in the mean AUROC score). This could be an artifact in the data; however, the same trend can be observed for all months, suggesting this time gap represents the greatest difference in seasonal ailments.



(a) ROC curves for the top 3 classifiers.

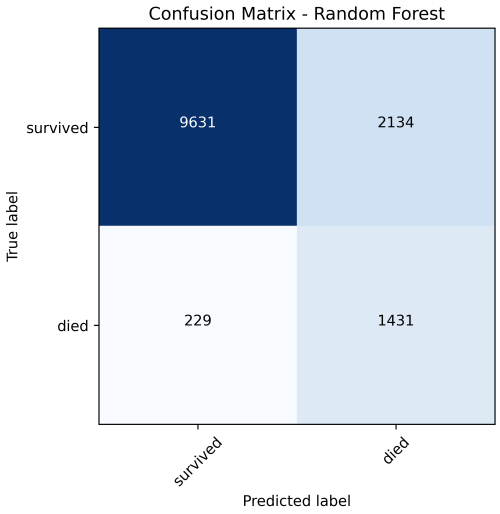


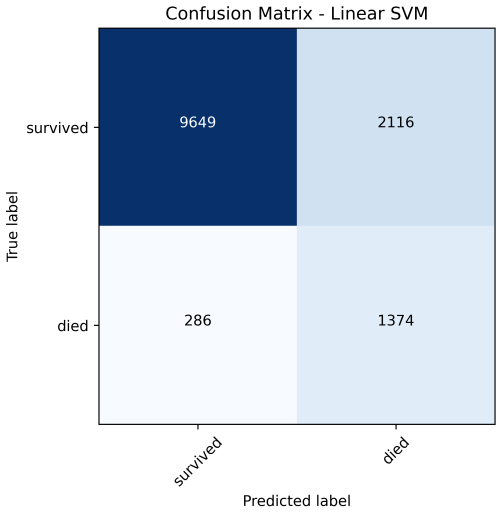
(b) ROC curves for the APACHE-II score.

Fig. 2: ROC curve comparison.

We then implemented a continuous retraining policy using a sliding window of two months. In this experiment we trained the random forest classifier on 2 months of data and then tested it against the next month of data, before sliding the window on a month and repeating the process. As can be seen from Figure 5, this results in much more consistent classifier performance across the year.

The results presented in Figure 5 show the mean of the performance metrics over 100 independent runs of the classifier training process as a solid line, and the standard deviation of the performance metrics as the shaded region. From this we can see that the AUROC is generally higher across most testing months than can be seen in Figure 4 - though there appears to be some impact as seasons change (e.g. when we test with data from June and train with data from April and May, we see a significant drop in predictive capability). As well as this, the variance between runs is significantly reduced, showing that our classifiers perform more consistently across different training runs.





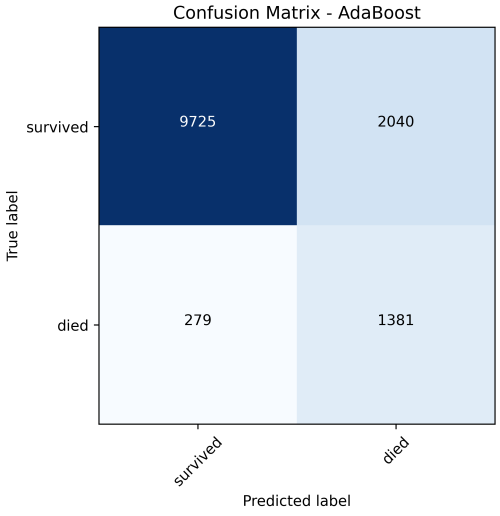


Fig. 3: Confusion plots for the top 3 classifiers.

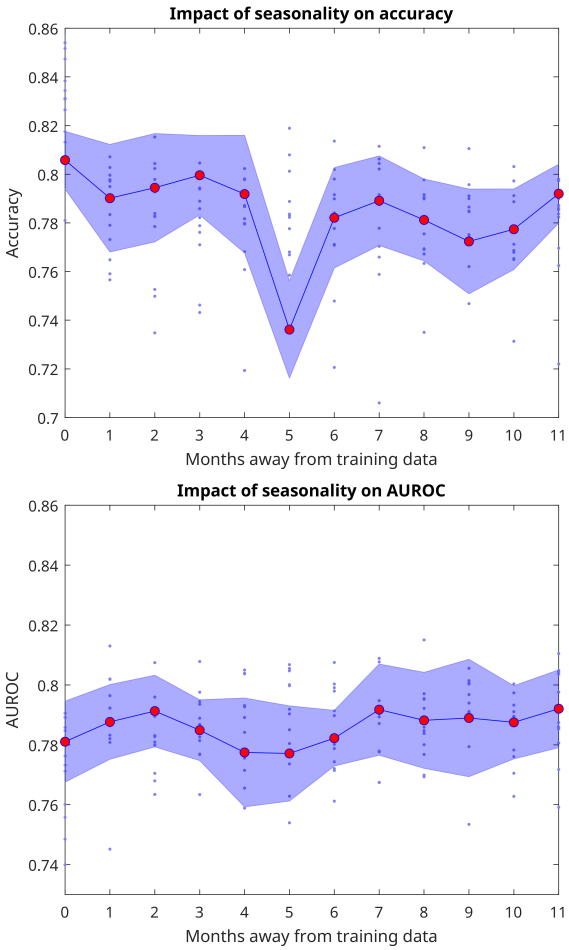


Fig. 4: The impact of seasonality on classifier performance.

2) Patient age: In this scenario the ICNARC data discussed in section III-A is partitioned based on patient age at admission into the ICU. Figure 6 shows the impact of ageing patient demographics on classifier performance where we can see that, as the patient cohort age group used for testing gets further away from that used in training (in this case patients aged less than 30), the classifier performance reduces dramatically. In particular, we can see that when testing on a patient population of over 80 years old, our trained classifier shows very poor predictive capability (with an AUROC of < 0.65). This may be because older adults have greater and more complex co-morbidities than the younger population and therefore mortality is more difficult to predict. This demonstrates the necessity of retraining and recalibration of predictive models for risk assessment as patient demographics change. As in section IV-B1 above, a random forest classifier was used as a strong baseline in these experiments.

We then implemented a continuous retraining policy by keeping the size of our training set static, but sampling from a wider range of ages to show the effectiveness of continuous adaption to changing patient demographics. Again, in this experiment we trained the ML classifier on the sampled data,

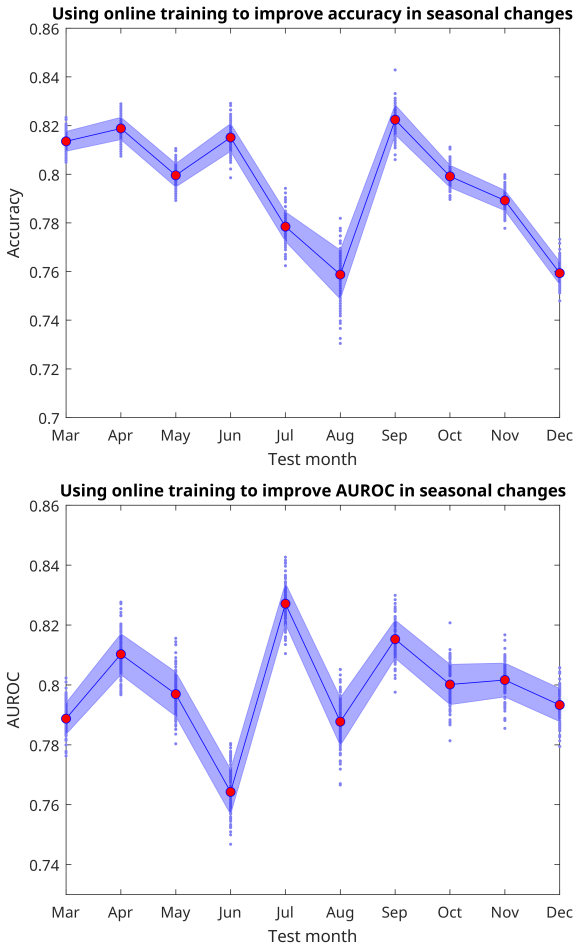


Fig. 5: Effects of online learning in dealing with seasonality.

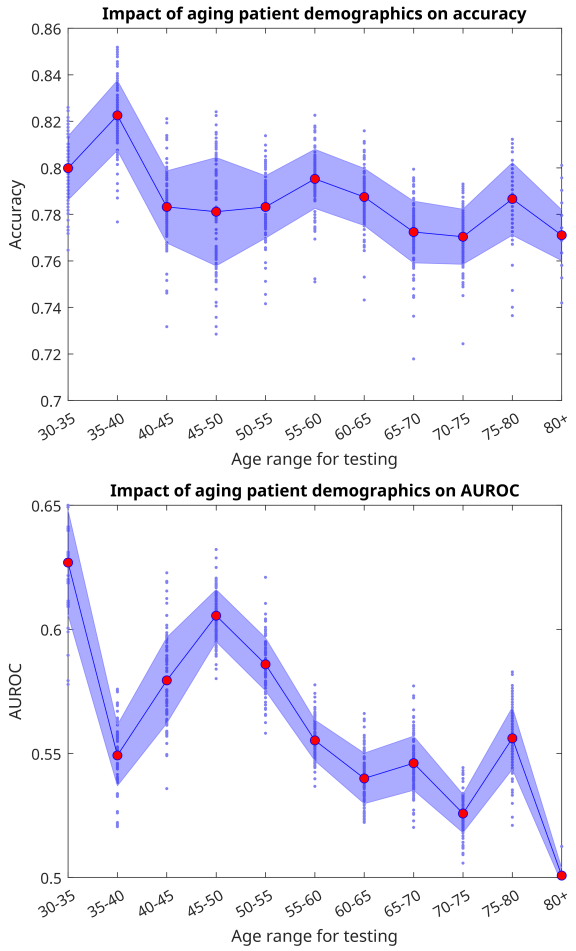


Fig. 6: The impact of aging patient demographics on classifier performance.

before testing against a patient cohort consisting of patients over 70 years of age. As can be seen from Figure 7, this results in better classifier performance (in terms of both accuracy and AUROC) particularly as the patient demographic group used in training includes samples closer to the testing demographic.

The results presented in Figure 7 show the mean of the performance metrics over 100 independent runs of the classifier training process as a solid line, and the standard deviation of the performance metrics as the shaded region. The training set was limited to 2000 patients randomly sampled (with stratification on outcome) from the age range shown on the x-axis.

We can see that as the age range used for training includes patients closer to the age range of the patient demographic used for testing the performance both in terms of accuracy and in terms of AUROC significantly improves, achieving a maximum AUROC of approximately 0.63 and maximum accuracy of 0.83 on the out of domain data (i.e., the testing patient cohort described above). As well as this, the variance between runs is also reduced, showing that our classifiers perform more consistently across different training runs.

V. Conclusions

The current study has indicated that ML can be used to accurately predict mortality in intensive care units. In this paper the ICNARC dataset from the North Middlesex University Hospital Cluster was used, which contains 29 features acquired within the first 24 hours of admission in to ICU. The most effective baseline models trained in this study included Random Forests, Linear SVMs, and the AdaBoost algorithm, achieving AUROC scores of 0.92, 0.91, and 0.91 respectively. These results indicate that the developed ML models are more effective than many current state-of-the-art techniques at predicting mortality on the patient cohort considered in this study.

The existing mortality risk prediction methods commonly used in ICU environments do not account for changes in medicine, patients reactions to intervention, changing patient demographics, and can not be calculated at admission to the ICU. This paper has presented the development of a ML pipeline that can be used to correctly identify and quantify the risk of mortality for patients admitted in to intensive care units and extended this pipeline to allow continuous retraining

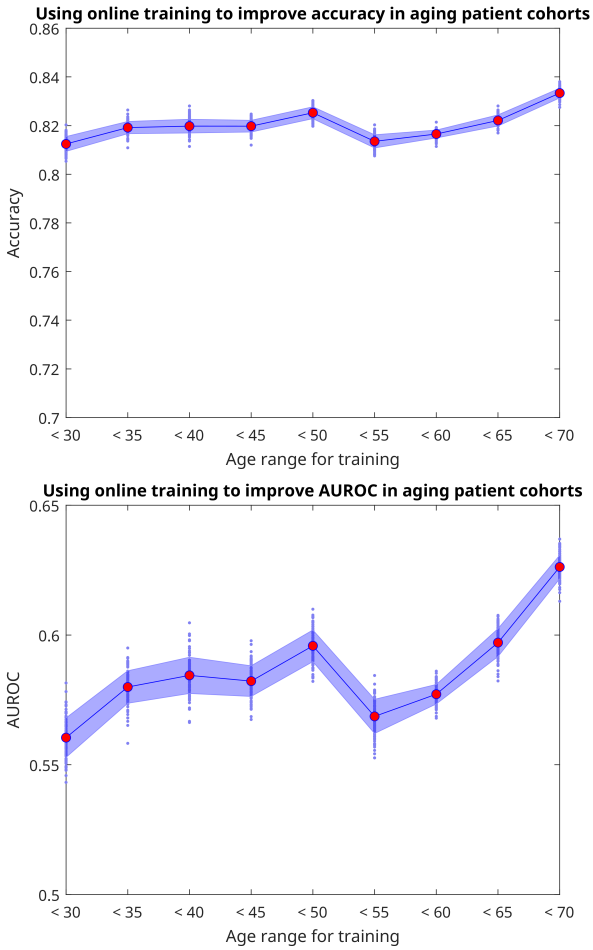


Fig. 7: Effects of online learning in dealing with ageing patient demographics.

via online learning. Results presented in section IV-B show how this can be used to improve classifier performance when patient demographics change. The results show that it is possible to train the model at different intervals for varying patient demographics to improve both the model accuracy and AUROC.

It is expected that further research and development into ML for mortality risk prediction will result in an online ML support tool that can be trained continuously with new data. This would make it suitable for application across many different patient cohorts, and allow the adaption of the model and methodology to be applied globally, as patient demographics evolve over time.

Funding Acknowledgement

This work was funded by Sheffield Hallam University's Graduate Teaching Assistant (GTA) PhD scholarship scheme.

References

[1] E. Devereux, "Nhs under 'huge strain' as staff sickness absences soar," https://www.nursingtimes.net/news/hospital/nhs-under-huge-strain-as-staff-sickness-absences-soar-23-12-2022/, accessed: 2023-04-10.

[2] J. C. Jentzer and X. Rossello, "Past, present, and future of mortality risk scores in the contemporary cardiac intensive care unit," *European Heart Journal Acute Cardiovascular Care*, vol. 10, no. 8, pp. 940–946, 2021.

[3] S. L. Barriere and S. F. Lowry, "An overview of mortality risk prediction in sepsis," *Critical care medicine*, vol. 23, no. 2, pp. 376–393, 1995.

[4] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (saps ii) based on a european/north american multicenter study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.

[5] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "Apache ii: a severity of disease classification system." *Critical care medicine*, vol. 13, no. 10, pp. 818–829, 1985.

[6] B. E. Keuning, T. Kaufmann, R. Wiersema, A. Granholm, V. Pettilä, M. H. Møller, C. F. Christiansen, J. Castela Forte, H. Snieder, F. Keus *et al.*, "Mortality prediction models in the adult critically ill: A scoping review," *Acta Anaesthesiologica Scandinavica*, vol. 64, no. 4, pp. 424–442, 2020.

[7] D. G. Arts, N. F. de Keizer, M. B. Vroom, and E. De Jonge, "Reliability and accuracy of sequential organ failure assessment (sofa) scoring," *Critical care medicine*, vol. 33, no. 9, pp. 1988–1993, 2005.

[8] A. Kramer, "Predictive mortality models are not like fine wine," *Critical Care*, vol. 9, p. 636, 2005.

[9] Y. Sakr, C. Krauss, A. C. Amaral, A. Rea-Neto, M. Specht, K. Reinhart, and G. Marx, "Comparison of the performance of saps ii, saps 3, apache ii, and their customized prognostic models in a surgical intensive care unit," *British journal of anaesthesia*, vol. 101, no. 6, pp. 798–803, 2008.

[10] C. C. H. Lew, G. J. Y. Wong, C. K. Tan, and M. Miller, "Performance of the acute physiology and chronic health evaluation ii (apache ii) in the prediction of hospital mortality in a mixed icu in singapore," *Proceedings of Singapore Healthcare*, vol. 28, no. 3, pp. 147–152, 2019.

[11] R. O. Deliberato, G. G. Escudero, L. Bulgarelli, A. S. Neto, S. Q. Ko, N. S. Campos, B. Saat, E. A. Júnior, F. S. Lopes, and A. E. Johnson, "SEVERITAS: an externally validated mortality prediction for critically ill patients in low and middle-income countries," *International journal of medical informatics*, vol. 131, p. 103959, 2019.

[12] T. Alves, A. Laender, A. Veloso, and N. Ziviani, "Dynamic prediction of icu mortality risk using domain adaptation," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1328–1336.

[13] M. Saeed, C. Lieu, G. Raber, and R. G. Mark, "Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring," in *Computers in cardiology*. IEEE, 2002, pp. 641–644.

[14] K. Yu, M. Zhang, T. Cui, and M. Hauskrecht, "Monitoring icu mortality risk with a long short-term memory recurrent neural network," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*. World Scientific, 2020, pp. 103–114.

[15] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[16] A. Shenfield, M. Rodrigues, H. Nooreldeen, and J. Moreno-Cuesta, "A novel hybrid differential evolution strategy applied to classifier design for mortality prediction in adult critical care admissions," in *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2017, pp. 1–7.

[17] J. Zhang and A. C. Sanderson, "Jade: adaptive differential evolution with optional external archive," *IEEE Transactions on evolutionary computation*, vol. 13, no. 5, pp. 945–958, 2009.

[18] S. Baker, W. Xiang, and I. Atkinson, "Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] P. Mishra, V. Varadharajan, U. Tupakula, and E. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 686–728, 2018.