Predicting flood inundation extent using remote sensing and machine learning techniques

D. Shrestha [a], D.E. Robertson [a], W. Jin [b] and C. Ticehurst [c]

[a] CSIRO, Environment, Melbourne, Australia
[b] CSIRO, Data61, Canberra, Australia
[c] CSIRO, Environment, Canberra, Australia
Email: durgalal.shrestha@csiro.au

Abstract: Flood inundation modelling and mapping are crucial for effectively managing river systems to minimize damage and protect ecosystems in floodplains and wetlands. This study develops a method that combines machine learning and remote sensing to predict flood inundation extent, which can be used to assess the impact of climate change on flood inundation. The method is applied to Macintyre River catchment in Australia, using Landsat-derived raster images from 1988 to 2020 and concurrent and lagged streamflow data as predictors. A range of ML techniques, including logistic regression, support vector machines, random forest, multi-layer perceptron and convolutional neural networks are employed. Three performance metrics namely accuracy, Heidke skill, and the area under the receiver operating characteristics curve are used to evaluate the performance of the ML models. The results show that the random forest model generally outperforms the other models, with accuracy well above 90% for balanced classes, while performance for the highly imbalanced class is worse than climatology. The performance of some ML models, particularly the convolutional neural networks and multi-layer perceptron, is sensitive to the randomization of initial parameter weights, so multiple runs are necessary for reliable results. The study concludes by suggesting future work to extend the method to simultaneously predict multiple pixels using a model that can produce multiple outputs or by incorporating spatial information in the model predictors. This would enable the assessment of the impact of climate change and various climate adaptation options on flood inundation extent.

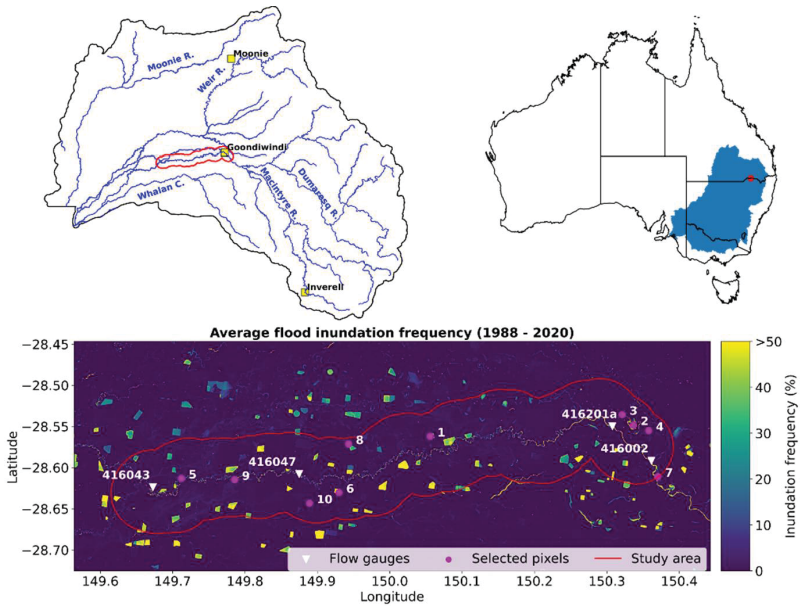

Figure 1. Location of study area

Keywords: *Flood inundation, machine learning, deep learning*

1. INTRODUCTION

Flood inundation modelling and mapping are critical for effective management of river systems, and can support minimizing loss of life and property, and preservation of healthy and resilient ecosystems in floodplains and wetlands. These models use advanced computational techniques to simulate flood events and predict the extent and depth of flooding in a given area. This information is then used to develop flood maps that show the areas at risk of flooding and the likely impact of floods on the environment, such as erosion, sediment deposition, habitat destruction and changes in nutrient availability. By providing detailed information on the likelihood and severity of flooding, flood inundation modelling and mapping enable policymakers and river managers to prioritize flood management strategies and identify areas where ecosystem protection and restoration efforts can be most effective.

Hydrodynamic models have been widely used to map flood inundation as they can realistically represent the physical process involved in modelling flood inundation (Schumann et al., 2016; Teng et al., 2017). However, these models are complex and computationally intensive, which can limit applications to analysis for large spatial domains, fine temporal and spatial resolutions, or long time series. There is an increasing need to develop cheap and accurate flood inundation mapping approaches. Remote sensing techniques have been utilized for producing flood inundation maps, as they provide basic data more cost-effectively and efficiently than ground-based methods (Whitehouse, 1989). These techniques analyse time series of different satellite images to produce inundation maps, showing how often an area has been inundated in the past. However, they are unable to provide information about the inundation extent given the current flow conditions or changes to inundation areas under climate change and different management scenarios.

Recently machine learning (ML) techniques have been used for flood inundation modelling. Most applications of ML generate surrogate models to emulate hydrodynamic models (e.g., Kabir et al., 2020; Xie et al., 2021). These applications have sought to characterize the dynamics of floods moving along a river valley to support emergency response. Climate change assessments do not require a detailed understanding of flood dynamics, rather, they require the ability to predict changes in inundation extent and frequency at seasonal or annual time scales. Shaeri Karimi et al. (2019) used a random forest to predict the flood inundation extent using coarse temporal resolution Landsat images describing the inundation. They selected 10,000 points randomly from the study area to create a training dataset that includes both spatial and temporal information on topographic, climatic and hydrological inputs. While this approach provides long records to train the machine learning model, the model is complex and can be computationally expensive to train.

In this paper, we describe a new method that combines ML and remote sensing techniques to predict flood inundation extent more quickly and with only hydrological inputs. We applied the method to predict flood inundation extent in the Macintyre River catchment in Australia. Sub-monthly Landsat-derived raster images mapping the inundation extent from 1988 to 2020 are used. Daily concurrent and lagged streamflow data from nearby gauges and recent flood inundation are used as predictors. We used a range of ML techniques, including logistic regression, support vector machines, random forest, multi-layer perceptron, and convolutional neural networks. Three performance metrics: accuracy, Heidke skill, and the area under the receiver operating characteristics curve are used to evaluate the performance of the ML models. We also compared the ML models to benchmark methods such as climatology and persistence. Additionally, we adapted a Bayesian approach to statistical significance testing to compare the performance of five ML models.

2. MATERIALS AND METHODS

2.1. Study area and data

The study area is located in the Macintyre River catchment in Australia. It spans approximately 110 km reach of Macintyre River, downstreams of Goondiwindi and covers 5 km on either side of the River (Figure 1). The flow in the Macintyre River is highly variable, with long-term daily flows at station 416002 estimated to be 520 ML. The long-term daily flow at station 416043 is approximately half of that at the most upstream site. The Macintyre River exhibits a summer-dominated annual flow pattern, with flood events more common between November and April.

Sub-monthly Landsat-derived raster images of the inundation extent were processed for a period from 1988 to 2020 using data from Digital Earth Australia (Geosci. Aust., 2023) and the method of Ticehurst et al. (2022). The majority of images have intervals of 8 (51.9%) and 16 (35.5%) days, while the rest of the images are available with intervals ranging from 1 day to 96 days, depending on cloud cover and the number of Landsat sensors operating at the time. The pixel size of the images is 25 by 25 m.

Since all machine learning models except CNN do not allow training and making predictions for all the pixels at one time, we have selected 10 random pixels with inundated frequencies between 20 to 50% in the period from 1988 to 2020 within a 5km buffer on either side of the Macintyre River. The lower threshold of 20% is selected to include wet events, while the higher threshold of 50% allows us to exclude permanent water bodies such as river channels, wetlands, and farm dams. Setting such criteria is critical for training ML models as it prevents the selection of pixels with highly unbalanced data sets, where one class (inundated or non-inundated) dominates over the other.

2.2. Methods

Machine learning methods

We used a range of machine learning techniques (Table 1) to predict whether a given pixel is inundated or not. Logistic regression (LGR) is a linear model commonly used for classification tasks. Support vector machines are supervised machine learning models that identify the optimum line or hyperplane in a multidimensional feature space to classify data where events falling on one side of the hyperplane are classified as positive and those on the other side are classified as negative. Random forest (RF) is a non-parametric algorithm that generates multiple decision trees, which are then aggregated to provide a final prediction. Artificial neural network (MLP), inspired by the biological nervous system, consists of interconnected processing units known as neurons and are widely used as universal function approximation due to their ability to represent both linear and complex non-linear relationships. Convolutional neural networks (CNN) are a specific type of neural networks that includes at least one convolutional layer that uses filters to perform convolution operations to extract features from the data.

LGR is simple and efficient model that can be easily interpreted and works well for linearly separable data. SVM generally works well for both linear and non-linear data by using different kernel functions and can handle high-dimensional data well. However, it can be computationally expensive for large datasets. RF works well for non-linear data and can handle high-dimensional data, but it can be also computational expensive for large datasets. MLP and CNN can handle non-linear data and learn complex relationship between features. However, both models require a lot of training data to generalize well and CNN can be computationally expensive for large networks and datasets.

Table 1. Machine learning models and hyperparameters

Models	Model complexity	Hyperparameters and values
LGR	Low	penally=L2, regularization C=1
SVM	Medium	kernel=radial basis, regularization C=1, epsilon=0.1
RF	Medium	n_estimators=100, max_depth=3, max_features=sqrt(n_features)
MLP	High	neurons=32, batch size=32, optimizer=Adam, epochs=5000
CNN	High	filter = [32,64,32], kernel=[3,3,3], neurons=1, batch size=32, optimizer=Adam, epochs=5000

Predictors selection

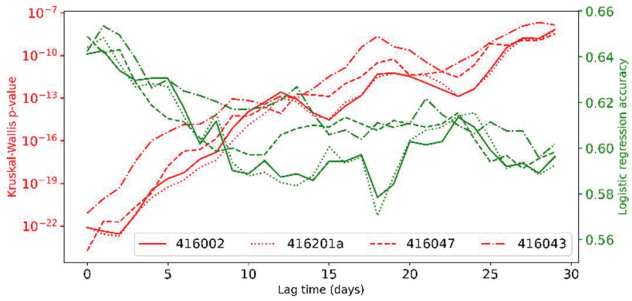

Figure 2. Kruskal-Wallis p-value and logistic regression accuracy showing the correlation between the flow data and flood inundation for pixel 5.

Predictor selection is a crucial step in ML. Due to the limited record of the flood inundation dataset, we aim to reduce model complexity by using a minimal number of input variables. We use streamflow data from nearby gauges as the magnitude of the flow predominantly determines whether a river will exceed its banks and cause inundation in surrounding areas. Specifically, we selected four gauges (416002, 416201a, 416047 and 416043). Because inundation is a cumulative effect of recent flows, we need to identify the appropriate lag time of the flow. While standard methods such as correlation analysis (May et al., 2008) were unsuitable due to the binary nature of the target variables, we employed Kruskal-Wallis test and logistic regression to estimate the correlation between lagged flow and flood inundation classes (0: non-inundation, 1: inundation). The low p-value of

Kruskal-Wallis test indicates a strong correlation between the flow and inundation class. The higher prediction accuracy from logistic regression shows that two variables (flow and inundation) are correlated. As shown in Figure 2, prediction accuracy declines as lag time increases up to 14 days, and remains relatively stable until a lag time of 22 days. Furthermore, the Kruskal-Wallis p-values demonstrate that the correlation between the flow and flood inundation weakens with increasing lag time, consistent with the logistic regression accuracy. Consequently, lag times of up to 14 days were selected for the study.

To train the ML models for a given target location, we used five different combinations of daily flow data from the nearest upstream and downstream streamflow gauges, referred to as 1us (one upstream), 1us1ds (one upstream and one downstream), 2us (two upstream), 2us1ds (two upstream and one downstream), and all sites (four gauges). In addition to streamflow gauges, we used recent flood inundation conditions as a predictor. For example, to classify pixel 5 on 1988/2/16 using the 1us predictor, we selected the flow data from 416047 gauge on 1988/2/15, 1988/2/14 ,…, 1988/2/2 and inundation condition on 1988/01/31. This resulted in a total of 15 input variables (14 lagged flows and one recent inundation condition). Similarly, for 1us1ds, 2us, 2us1ds, and allsites predictors, 29, 29, 43, and 57 input variables were used, respectively.

2.3. Validation

The performance of the machine learning models was evaluated by splitting the dataset into training (70%), and test sets (30%). Three models, namely LGR, SVM, and RF, were trained on the training data. For MLP and CNN, the training data was further divided into training (80%) and cross-validation (20%) sets. The cross-validation dataset was used for these two models to prevent over-fitting of the models during training. To quantify the classification error, three matrices were used, namely accuracy, Heidke skill score (HSS), and area under the receiver operating characteristics (AUC).

Accuracy measures the fraction of the total events ($N_{hits} + N_{correct\_negatives}$) correctly classified, N_{hits} is the number of times the given pixel is correctly classified as inundated, $N_{correct\_negatives}$ the number of times the given pixel is correctly classified as non-inundated. HSS measures the prediction accuracy relative to that of random chance (often referred to as expected correct) and is defined as:

$$HSS = \frac{accuracy - expected\_correct}{1 - expected\_correct},\ expected\_correct = \frac{1}{N^2}(N_{inundated}\hat{N}_{inundated} + N_{non-inundated}\hat{N}_{non-inundated}) \quad (1)$$

where, $N_{inundated}$ is the number of inundated observations and $N_{non-inundated}$ is the number of non-inundated observations, $\hat{N}_{inundated}$ is the number of inundated events that are correctly classified and $\hat{N}_{non-inundated}$ is the number of non-inundated events correctly classified. The HSS score ranges from - ∞ to 1, where a value of HSS > 0 indicates better performance than the random chance. A value of HSS = 1 indicates a perfect classification, while a value of HSS < 0 indicates worse performance than the random chance. For highly imbalanced datasets, random chance cannot be used as a reference forecast, and instead we used climatology prediction as a reference prediction. For instance, if $N_{inundated} = 0$ *and* $\hat{N}_{inundated} = 0$ then $expected\_correct = 1$ and consequently HSS cannot be computed.

Receiver operating characteristics (ROC) measures the discrimination between two alternative outcomes or classes (e.g., inundated and non-inundated). The ROC curve plots the hit rate ($N_{hits} / N_{imundated}$) against the false alarm rate ($N_{false\_alarms} \setminus N_{non-inundated}$) for a range of decision probability thresholds and can be summarized by computing the area (AUC) under the ROC curve. N_{false_alarms} is the number of times the given pixel is wrongly classified as inundated while it is non-inundated. A perfect classification will have an AUC equal to 1.0 while an unskilled classification will have an AUC of 0.5 or less.

Due to the internal randomness of some learning algorithms, the performance of a model can vary when it's run again on the same dataset. For instance, MLP is initialized with random weights, so if it's trained again with different initializations, it may produce different results even if the training data remain the same. To embrace the stochastic nature of machine learning training, the model is trained on the same dataset 100 times and the performance of all 100 runs is reported with box-whisker plots. The box-whisker plots display the variations in the model's performance but do not provide certainty of the difference between models. Therefore, a significance test is conducted to statically compare the performance of the machine learning models. Bayesian estimation is used to calculate the probability of one model being better than another. The posterior of the mean parameter μ can be modelled by a Student's t-distribution (Benavoli et al., 2017):

$$St(\mu, r-1, \bar{x}, (\frac{1}{r} + \frac{N_{test}}{N_{train}})\hat{\sigma}^2) \qquad (2)$$

where, r is the number of repeats (model runs), $\bar{x}$ is the mean difference in the performance scores, N_{test}, N_{train} are the number of samples used for testing and training, respectively, $\hat{\sigma}^2$ is the variance in the difference of the performance scores. The probability that one model is better than the other is determined by the area under the curve of the posterior distribution from zero to infinity. Similarly, the probability that one model is worse than the other is determined by the area under the curve of the posterior distribution from minus infinity to zero. The probability that both models perform equivalently is computed by finding the area under the curve of the posterior over the interval [-0.01, 0.01].

In addition to five ML models, two reference predictions: climatology and persistence were used. Climatology predicts flood inundation conditions in the test dataset based on the seasonal climatology of the flood inundation conditions in the training dataset. Persistence uses the flood condition of the recent event to predict the next event. Any machine learning models that perform worse than these benchmark forecasts may have limited or no practical value.

3. RESULTS

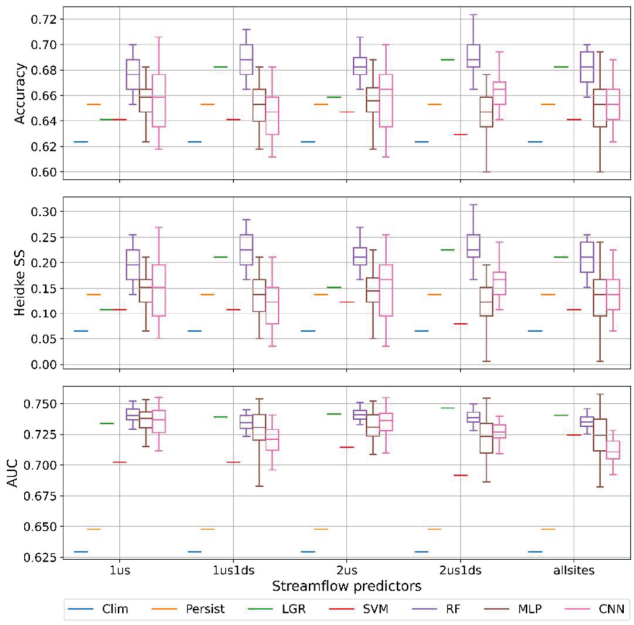

Figure 3. Performance comparison of different machine learning models for pixel 5. Each box represents the distribution of model performance metrics, with the central line indicating the median, and the bottom and top edges of the box representing the 25th and 75th percentiles, respectively. The bottom and top whiskers extend to the 10th and 90th percentiles, respectively.

In Figure 3, we present the performance metrics of different models in the test dataset for pixel 5 (see Figure 1). RF outperforms all other models when the daily flow from two or fewer gauges is selected as the predictor. However, when the daily flow from more than two gauges is used, LGR performs better than RF. SVM exhibits the worst performance across all predictor combinations, and it even underperforms persistence prediction in terms of accuracy and HSS. Compared to the climatology prediction, LGR and SVM, as well as the median of RF, MLP, and CNN, perform better in terms of accuracy and HSS. The AUC score of all ML models is superior to both reference predictions. It should be noted that the performance of RF, MLP, and CNN models shows significant variation between runs on the same dataset due to several factors, including intrinsic randomness in the model-building process, a higher number of parameters that need to be trained relative to the training dataset size, suboptimal model structure and hyperparameters, and poor data splitting for training, validation, and testing (May et al., 2010). For instance, CNN was trained on only 316 samples to estimate 12,801 weights and biases parameters. When comparing streamflow predictors, the performance varies across the models. 2us1ds predictors produced better performance matrices for LGR and RF models, whereas 2us predictors resulted in better performance matrices for SVM, MLP and CNN models. However, it is worth noting that the difference in performance among models when using different predictors is relatively small compared to the magnitude of the performance variance observed in MLP and CNN models trained multiple times on the same dataset. Figure 4 depicts the performance of ML models using best predictors over all ten pixels in the test data. As we observe, all ML models outperform the climatology prediction except for very dry pixels. Since, the climatology prediction is already incorporated into the HSS, we excluded them from this comparison as it is not a stringent reference prediction. It is important to note that the best predictors for one pixel may not be the best predictors for another due to the relative location of the pixels and streamflow gauges. Therefore, in this study, the best predictors are defined as those with the highest median AUC value after pooling them from all five ML models. As expected, the

performance of ML models display variability across the different pixels. For pixels with a with a balanced class, most of the ML models demonstrate very high performance. For instance, pixels 1, 2, 6, and 10, which have balanced classes show accuracy well above 90%. The accuracy of most of the ML models is also well above 90% for pixels 4 and 9, which have highly imbalanced classes. However, for these pixels, HSS is negative, indicating that the performance of the ML models is worse than that of climatology prediction. Notably, for very dry pixel 9, where the percentage of non-inundated class is around 2%, it is possible to achieve 98% accuracy by predicting non-inundated class for every event (Finley affair, Murphy, 1996).

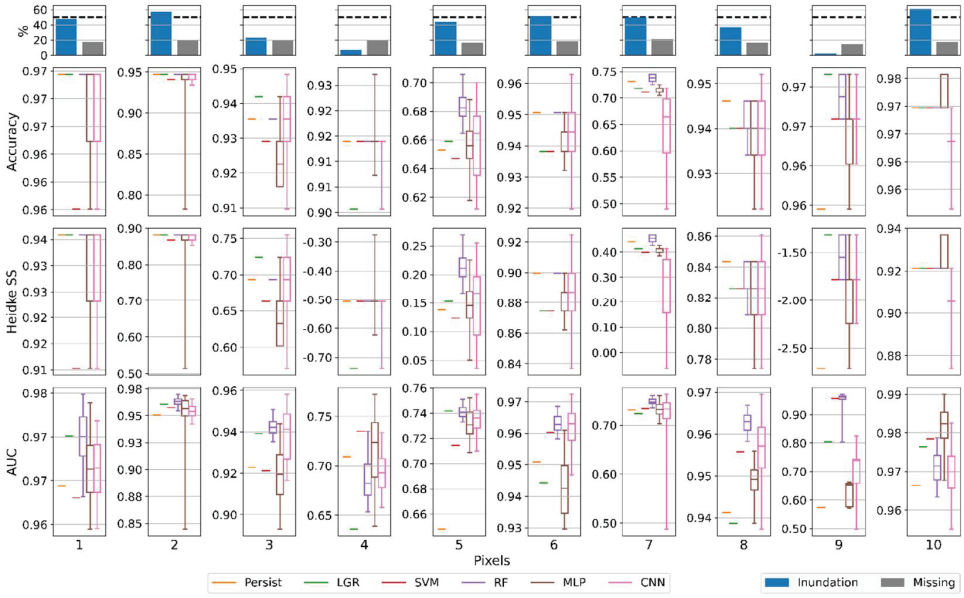

Figure 4. Comparison of model performance. For each pixel, the predictor with the best performance was selected. The top figure panel show the percentage of inundation and missing Landsat dataset.

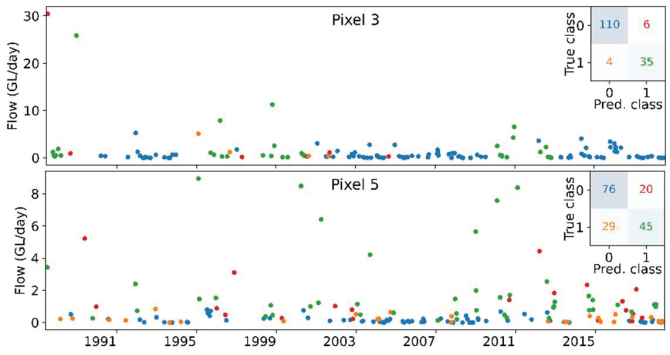

Figure 5. Example of RF predictions on pixels 3 and 5 using single us and ds predictors. Class 0 indicates non-inundation and class 1 indicates inundation. The confusion matrix is shown in the inset. The markers in the time series are coloured according to the confusion matrix, i.e., green is for

Therefore, caution must be exercised while using accuracy scores, particularly for imbalanced datasets. Although pixels 5 and 7 have balanced classes, the performance of ML models is not as good. Both pixels are very close to the river channel and are susceptible to frequent flooding, which is not well captured by the ML models. Pixel 7 does not have upstream flow gauges in this study, making it challenging to model using only downstream flow gauges. Despite having less balanced classes, the performance of the ML models for pixels 3 and 8 is reasonably good. In Figure 5, time series of RF predictions and upstream flow for two pixels 3 and 5 are presented. Correctly classified flood inundation conditions are denoted by blue and green colors, while incorrectly classified conditions are represented by red and orange colors. For the imbalanced dataset of pixel 3, RF was able to accurately classify most events as flood or non-flood. The flood inundation condition was found to be closely related to the magnitude of upstream flow. Interestingly, the highest flow event (the first point in pixel 3) was classified by RF as inundation, while the true class was non-inundation. This may be a classification error from the Landsat. Similarly in pixel 5, two events where the flow exceeded 4 GL/day were classified as non-inundation by Landsat, but they are likely to be actual inundation events.

Table 2 reports the results of a statistical significance test, conducted to compare the performance of the five machine learning models. The test was based on 100 runs for each pixel, as shown in Figure 4. The values in

the upper diagonal of the table represent the probability that the model in the row index performs better than the model in the column index, while values in parentheses represent the probability of models being practically equivalent. In contrast, the values in the lower diagonal indicate the probability that the model in the row index performs worse than the model in the column index. For instance, the RF model has a 0.31 probability of performing better than the SVM model, a 0.43 probability of being practically equivalent, and a 0.27 probability of performing worse than the SVM model. The statistical significance test results indicate that RF and SVM models perform better than CNN, LGR, and MLP models. Out of the five models, the LGR and MLP models perform the worst.

Table 2. Statistical significance test

Models	RF	SVM	CNN	LGR	MLP
RF	-	0.31 (0.43)	0.66 (0.12)	0.70 (0.19)	0.66 (0.10)
SVM	0.27	-	0.64 (0.12)	0.66 (0.18)	0.65 (0.10)
CNN	0.22	0.24	-	0.32 (0.21)	0.46 (0.20)
LGR	0.11	0.17	0.47	-	0.52 (0.17)
MLP	0.23	0.25	0.34	0.31	-

4. CONCLUSIONS AND RECOMMENDATIONS

This study presents a new methodology that combines machine learning (ML) and remote sensing to predict flood inundation extent, which can be useful for assessing the impact of climate change on flood inundation. Several ML techniques are employed, and their performance was assessed using three metrics: accuracy, Heidke skill, and the area under the receiver operating characteristics curve. The results showed that the models' performance varied among pixels, with RF generally performing better than the other methods. For pixels with balanced classes, accuracy exceeded 90%, whereas performance was worse than climatology for highly imbalanced classes. Additionally, we found that some of the models, particularly CNN and MLP, were sensitive to the randomization of the initial parameter weights, making it necessary to use multiple runs to obtain reliable results.

The hyperparameters for the models in this study were selected based on a review of the literature and to some extent through trial and error. It is anticipated that model performance could be improved by conducting proper hyperparameter tuning. Future research will aim to extend the methodology to simultaneously predict multiple pixels using a model capable of producing multiple outputs (e.g., CNN), or by integrating spatial information into the model predictors (e.g., DEM, distance to rivers, etc.). This approach has the potential to assess the impact of climate change and different climate adaptation strategies on flood inundation extents, thus providing valuable insights for flood risk management and adaptation planning.

ACKNOWLEDGEMENTS

This work is funded by Digital Water and Landscapes strategic project of CSIRO Environment.

REFERENCES

Geosciense Australia, G., 2023. DEA Surface Reflectance, https://www.dea.ga.gov.au/products/dea-surface-reflectance. Accessed 16 March 2023.

Benavoli, A., Corani, G., Demšar, J., Zaffalon, M., 2017. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. The Journal of Machine Learning Research, 18(1): 2653-2688.

Kabir, S. et al., 2020. A deep convolutional neural network model for rapid prediction of fluvial flood inundation. J. Hydrol., 590.

May, R.J., Maier, H.R., Dandy, G.C., Fernando, T.G., 2008. Non-linear variable selection for artificial neural networks using partial mutual information. Environmental Modelling & Software, 23: 1312-1326.

Murphy, A.H., 1996. The Finley affair: A signal event in the history of forecast verification. Wea. Forecasting, 11(1): 3-20.

Schumann, G.J.-P. et al., 2016. Rethinking flood hazard at the global scale. Geophysical Research Letters, 43(19): 10249-10256.

Shaeri Karimi, S., Saintilan, N., Wen, L., Valavi, R., 2019. Application of Machine Learning to Model Wetland Inundation Patterns Across a Large Semiarid Floodplain. Water Resour. Res., 55(11): 8765-8778.

Teng, J. et al., 2017. Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. Environmental Modelling & Software, 90: 201-216.

Ticehurst, C.J., Teng, J., Sengupta, A., 2022. Development of a multi-index method based on Landsat reflectance data to map open water in a complex environment. Remote Sensing, 14, 1158.

Whitehouse, G. (Ed.), 1989. Flood monitoring and floodplain studies. Remote Sensing in Hydrological and Agrometeorological Applications. COSSA Publications: Canberra.

Xie, S. et al., 2021. Artificial neural network based hybrid modeling approach for flood inundation modeling. J. Hydrol., 592: 125605.