Deep Learning in Estimation of Fruit Attributes Using Near Infrared Spectroscopy

Jeremy Walsh, Master of Research

INSTITUTE FOR FUTURE FARMING SYSTEMS

Presented by Dr. Zhenglin Wang

RESEARCH WITH IMPACT

Industry Adoption

Felix Instruments F750 Produce Quality Meter

- Used through the Australian mango value chain to achieve higher eating quality fruit
- Typically, PLSR models are not robust across varieties, physiological stages, growing conditions, seasons, and individual sensor instruments
- ANN models now used by Felix Instruments
- Could deep learning further improve model robustness and prediction accuracy?

Review Article

Journal of Near Infrared Spectroscopy
2022, Vol. 30(1) 3–17

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09670335211057235
journals.sagepub.com/home/jns
SAGE

Review: The evolution of chemometrics coupled with near infrared spectroscopy for fruit quality evaluation

Nicholas T Anderson[1] and Kerry B Walsh[1]

Review Article

Journal of Near Infrared Spectroscopy
2023, Vol. 31(3) 109–125

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09670335231173140
journals.sagepub.com/home/jns
Sage

Review: The evolution of chemometrics coupled with near infrared spectroscopy for fruit quality evaluation. II. The rise of convolutional neural networks

Jeremy Walsh, Arjun Neupane, Anand Koirala, Michael Li and Nicholas Anderson

Progression of techniques over the past three decades

- From MLR to PLSR
- More recently emergence of SVMs, ANNs

The Rise of Convolutional Neural Networks

- Improvements in computing power and availability of large training datasets have supported a trend to deep learning techniques

RESEARCH WITH **IMPACT**

Mendeley Data

Mango DMC and spectra Anderson et al. 2020

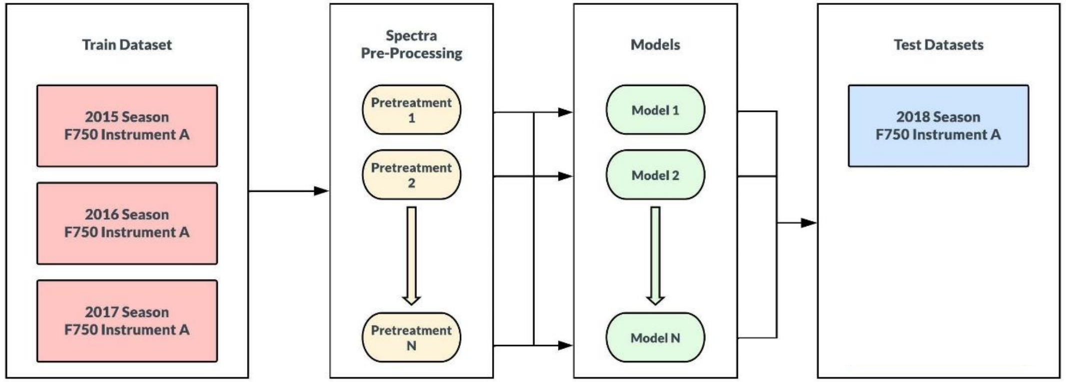
Published: 27 August 2020 | Version 2 | DOI: 10.17632/46htwnp833.2
Contributors: Nicholas Anderson, Kerry Walsh, Phul Subedi

Largest publicly available fruit NIR spectroscopy dataset

- 11,834 spectra samples from one F750 Produce Quality Meter
- 4,685 mangos used
- 112 unique populations
- Spans four growing seasons
- Multiple cultivars and growing locations

Featured in 10 publications and counting across multiple groups

- First three seasons used for training
- Fourth season used for independent validation
- Allows direct comparison of RSMEP of global modelling techniques



RESEARCH WITH IMPACT

CQUniversity AUSTRALIA

Publication (Wavelength Range)	Outlier Removal	Data Pre-Processing	Model	RMSEP
Anderson, Walsh et al. [87] **(684 - 990 nm)**	No additional	MC + SAVGOL (deriv=2, window=17, poly=2)	PLSR	1.014
			ANN	0.892
Anderson, Walsh et al. [89] **(684 - 990 nm)**	No additional	MC + SAVGOL (deriv=2, window=17, poly=2)	Gaussian Process Regression (GPR)	0.898
			Memory Based Learner (MBL)	0.903
			Hone Create Stacked Ensemble	0.85
Mishra and Passos [77] **(684 - 990 nm)**	No additional	Raw absorbance spectra	PLSR	1.06
	No additional	Data augmented by stacking: - Raw spectra - SNV - SAVGOL (deriv=1, window=13, poly=2) - SAVGOL (deriv=2, window=13, poly=2) - SNV + SAVGOL (deriv=1, window=13, poly=2) - SNV + SAVGOL (deriv=2, window=13, poly=2)	PLSR	1.03
	Removed using Hotelling's T2 and Q stats with PLSR decomp Train set 9914 (2015-2017 seasons) Test set 1,413 (2018 season)		PLSR	0.99 0.95*
			Cui and Fearn [28] 1D-CNN (kernal size = 21, batch size = 128) Involves decision making based on user experience to choose best hyper-parameters	0.79 0.75*
Yang, Luo et al. [76] **(684 - 990 nm)**	No additional	Raw spectra	1D CNN (3 Convolutional layers), with transfer learning applied with x% of season 4 data	1** (x=5%) 0.71*** (x=20%)
Mishra and Passos [81] **(684 - 990 nm)**	Removed using Hotelling's T2 and Q stats	Raw spectra	Cui and Fearn [28] 1D-CNN with transfer learning applied with 60% of season 4 data	0.58**

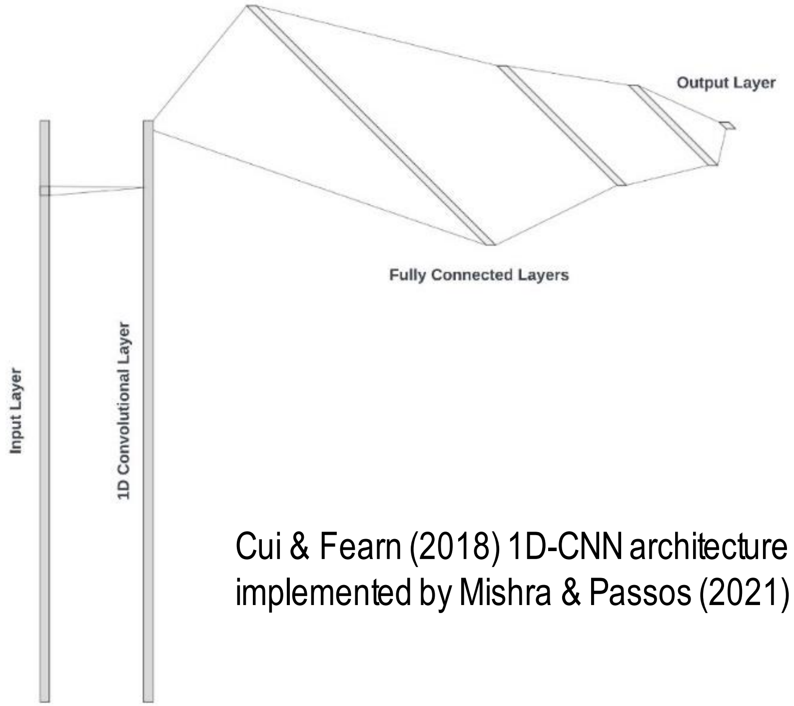
*Additonal 'outliers' removed from original test set

**Original test set data used to tune model

1D CNN 0.79% FW

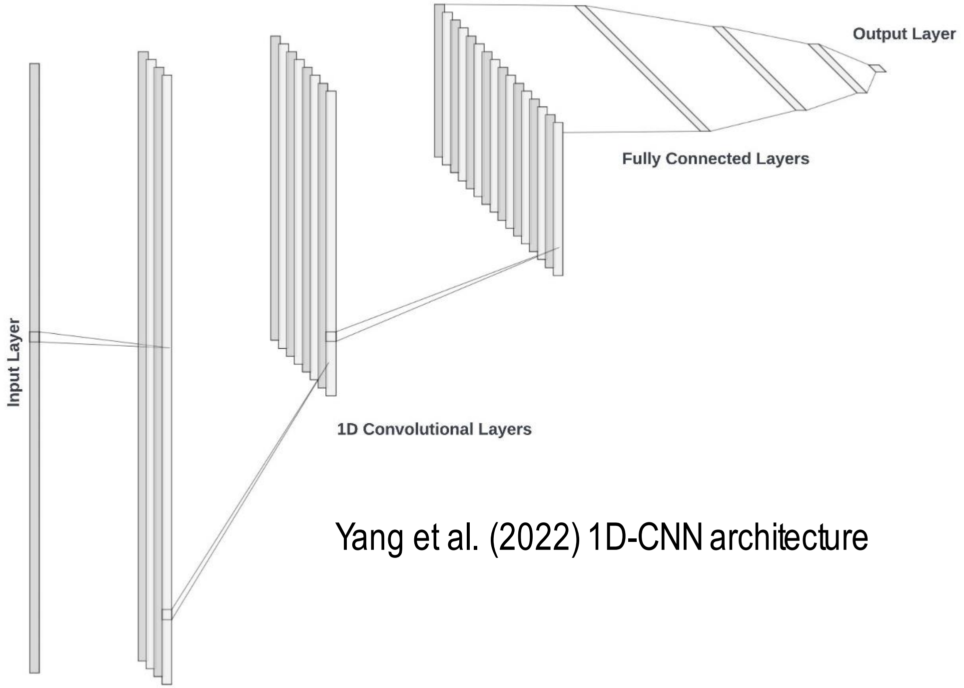
Deeper wasn't always better

RMSEP of 0.76



Cui & Fearn (2018) 1D-CNN architecture implemented by Mishra & Passos (2021)

RMSEP of 0.79 (10% test set used for transfer learning)
RMSEP of 1.0 (with 5%)

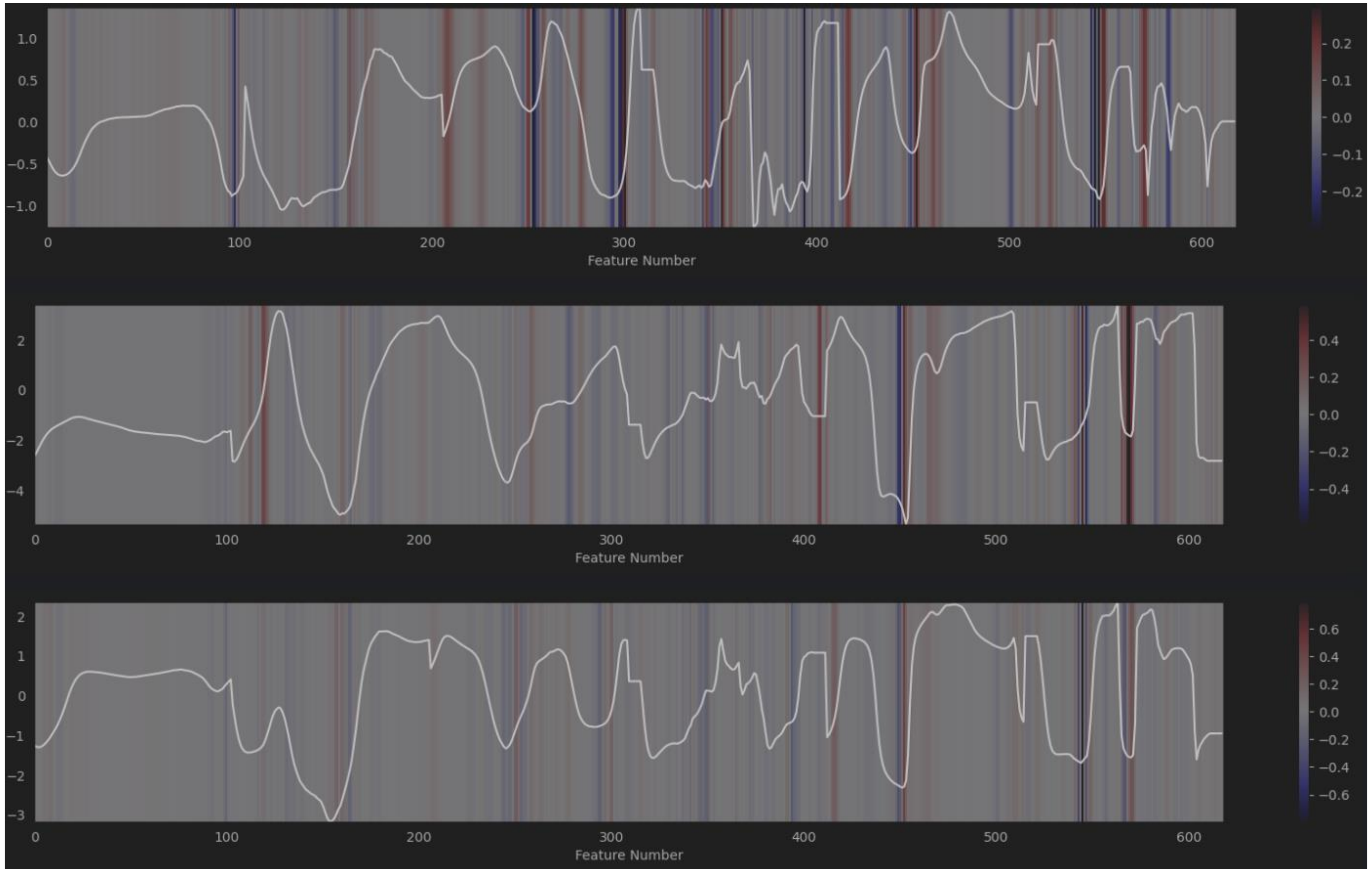


Yang et al. (2022) 1D-CNN architecture

Interpreting the “Black Box”

Gradient-weighted Class Activation Mapping (GradCAM)

- Mishra & Passos (2021) 1D-CNN model used augmented data input by combining an ensemble of different pre-treatments
- GradCAM output is specific to each test sample, this shows three randomly selected spectra from the independent test below



RESEARCH WITH IMPACT

ELSEVIER

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

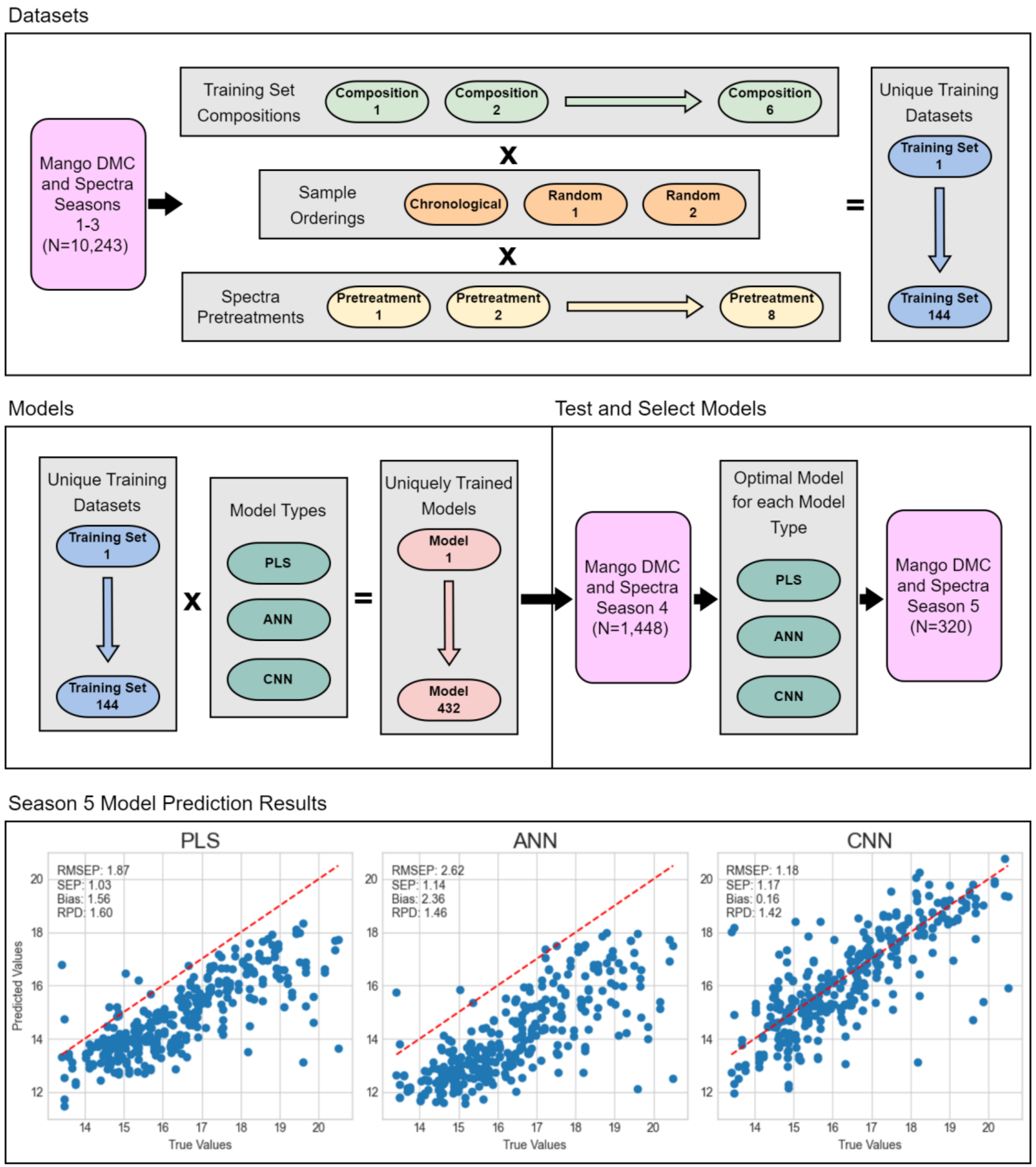
Volume 311, 15 April 2024, 124003

Evaluation of 1D convolutional neural network in estimation of mango dry matter content

Jeremy Walsh, Arjun Neupane, Michael Li

Highlights

- Comparison of model types on a 'level playing field' to deconfound results across multiple studies.
- Validation of CNN model superiority, outperforming PLS/ANN in mango DMC prediction.
- Benchmark RMSEP of 0.77%FW achieved, using data augmentation.
- Recommendations for quantifying neural network sensitivity to random seeds.
- Study provides insights for practical integration of CNN into portable instrumentation.

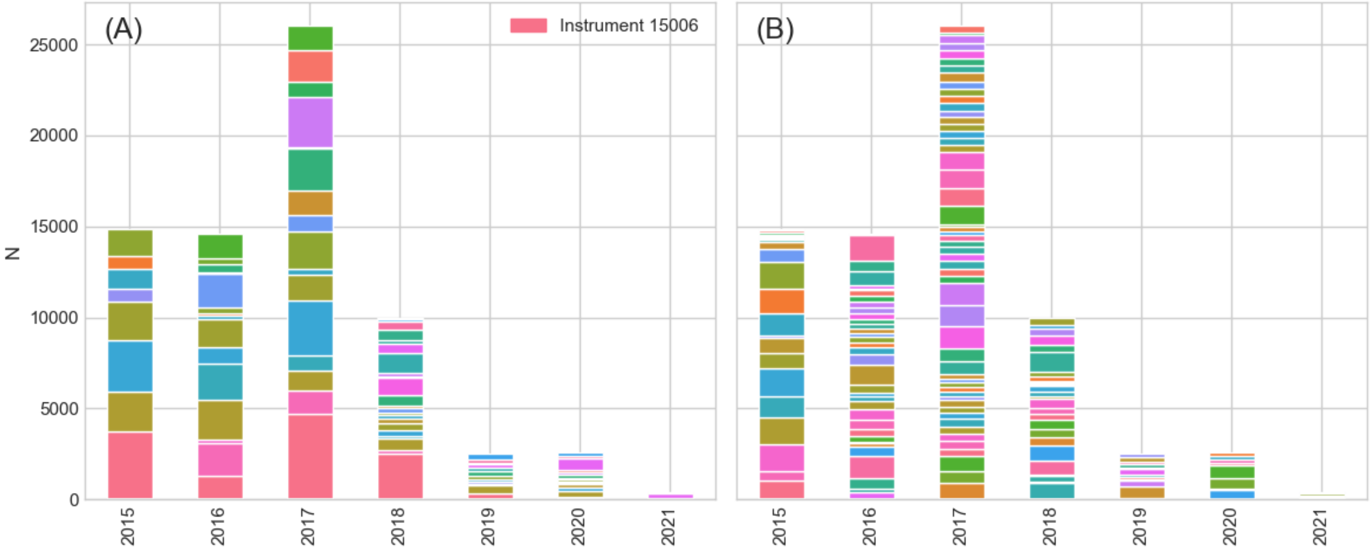


Datasets
Mango DMC and Spectra Seasons 1-3 (N=10,243)
Training Set Compositions
Composition 1
Composition 2
Composition 6
x
Sample Orderings
Chronological
Random 1
Random 2
x
Spectra Pretreatments
Pretreatment 1
Pretreatment 2
Pretreatment 8
=
Unique Training Datasets
Training Set 1
Training Set 144
Models
Unique Training Datasets
Training Set 1
Training Set 144
x
Model Types
PLS
ANN
CNN
=
Uniquely Trained Models
Model 1
Model 432
Test and Select Models
Mango DMC and Spectra Season 4 (N=1,448)
Optimal Model for each Model Type
PLS
ANN
CNN
Mango DMC and Spectra Season 5 (N=320)
Season 5 Model Prediction Results
PLS
RMSEP: 1.87
SEP: 1.03
Bias: 1.56
RPD: 1.60
ANN
RMSEP: 2.62
SEP: 1.14
Bias: 2.36
RPD: 1.46
CNN
RMSEP: 1.18
SEP: 1.17
Bias: 0.16
RPD: 1.42
Predicted Values
True Values

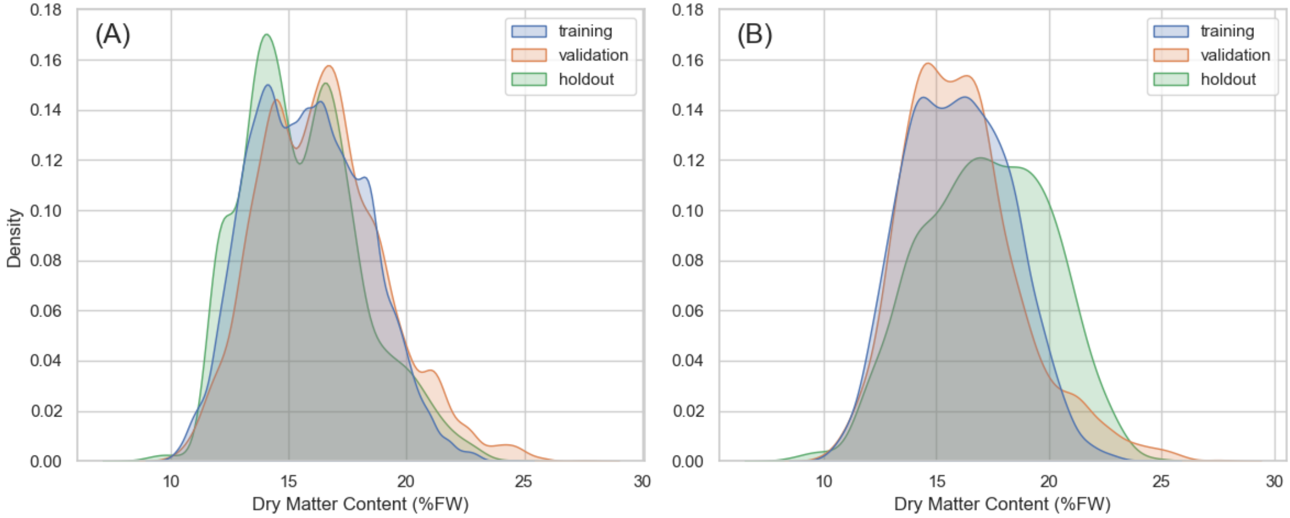
Mendeley Data

Mango DMC and NIR spectra

Published: 8 May 2024 | Version 5 | DOI: 10.17632/46htwnp833.5
Contributors: Nicholas Anderson, Kerry Walsh, Phul Subedi, Jeremy Walsh



Number of samples per season, colour coded by instrument (A) and population (B).



Distribution of DMC reference values for each partition of the dataset for all spectra samples (A) and for unique reference values (B).

Updated to 85,000+ spectra records

- Now spans 7 growing seasons
- 199 unique fruit populations
- 31 distinct instruments
- 10,560 unique refence values

RESEARCH WITH IMPACT

CQUniversity AUSTRALIA

Optimisation of deep learning models

» Benchmarking the previously reported optimal PLS, ANN and CNN models developed for the single-instrument mango DMC dataset on the new multi-instrument dataset.

» Optimisation of the wavelength range and preprocessing techniques used in context of a deep learning model rather than a PLS model.

» Optimisation of the deep learning architecture of a CNN model, in context of number of convolutional layers, convolutional layer kernels, dense layers, dense layer units, dropout rate and hyperparameters of the model.

» Optimisation of ANN and PLS models in context of wavelength range, preprocessing techniques, model architecture and hyperparameters.

» A comparison of the optimal PLS, ANN and CNN models on an independent (separate seasons) holdout set.

RMSEP (%FW)

	cal holdout	next two seasons + instruments
PLS	1.19	0.93
ANN	0.96	1.24
CNN	0.96	1.37

- The CNN model has over-fitted to the training set
- Where is the new variation coming from?

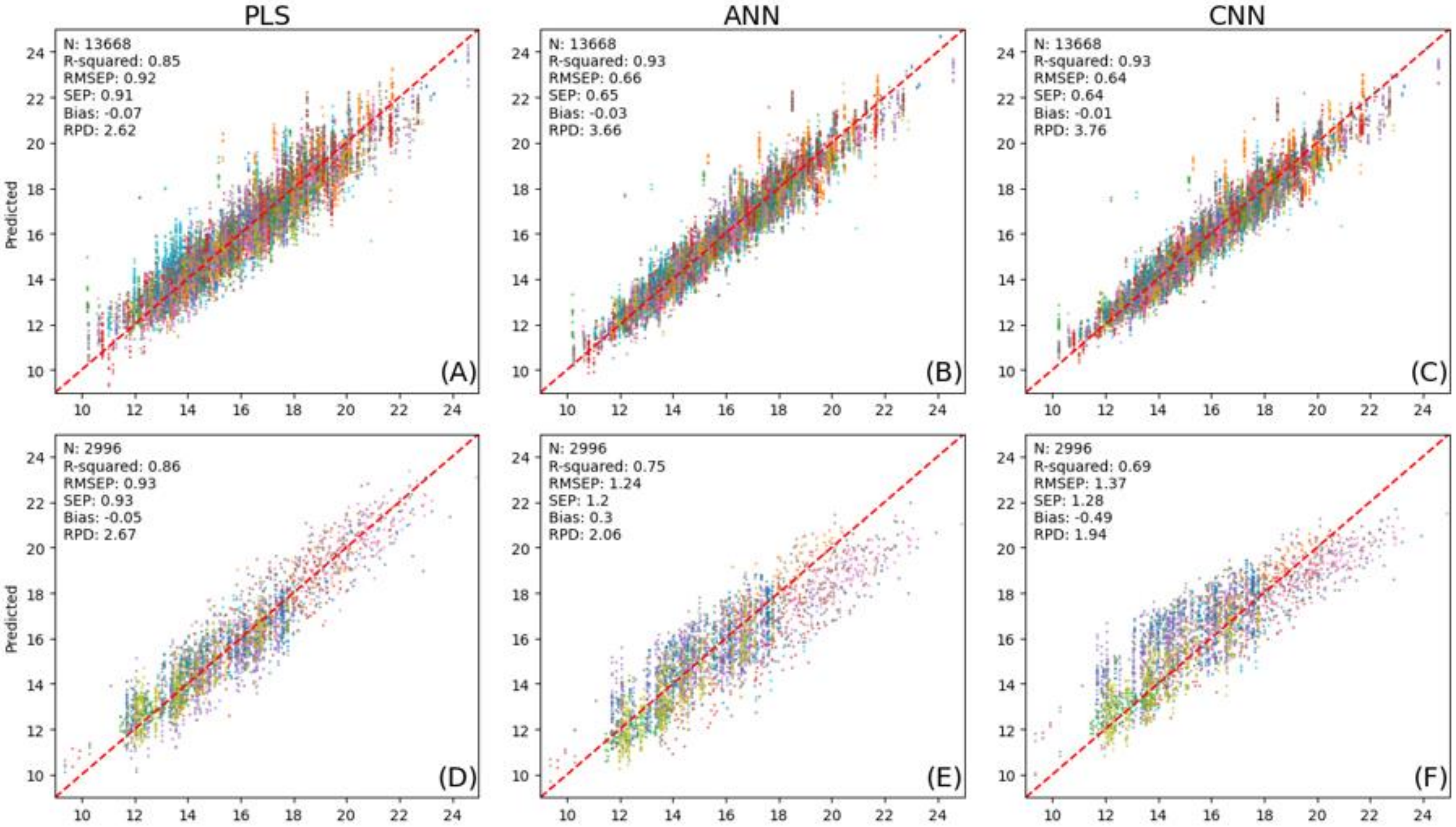


Figure 4.8: Plot of predicted against reference DMC (%FW) values for the combined tuning set (A, B, C) and holdout set (D, E, F), based on optimised PLS (A, D), ANN (B, E) and CNN (C, F) models.

The plotted prediction results are for the ANN and CNN models that performed the best on the tuning set, of the ten trained models. Samples are colour coded by population.

Future research

» Rigorous Model Evaluation Techniques.
» Enhanced Model Optimisation.
» Broader and More Representative Datasets.
» Utilisation of Cloud Computing.
» Investigation of Spectral Range and Pretreatments.
» Practical Implementation Considerations.