We have added several references to the following paragraphs:

Model evaluation:

Maybe an example of K-fold cross validation. And explain the intention of K-fold CV.

Line 61: why complex model lead to high evaluation variance

Line 63: trade off between bias-variance

Haque 2023: when it tried to compare the result 99.1% with others, it worth keeping in mind that how the accuracy is estimated (10-fold in this example) and the sample-feature ratio. For example, it compared with another study Sibiya that have 70%/30 split (nearly 10-fold) with 100 images that report 92.85% as final accuracy. Such comparision could be misleading as they are compared using different estimator and sample size, resulting in incomparable of estimated performance in terms of bias and variance.

Model selection:

Feature selection: Zhang 2019 – select 6 out of 415-1008nm (470 bands) using successive projections algoritm (SPA) based on cliinearity

Hyperparameter tuning

* Zhang 2019 use 5 fold to search number of tree (400, 600), number of spectral band (10-30). For SVM, C in 1, 10, 100, 1000 and kernel of RBF and linear is selected
* Becker – use Grid search to find the betst regularization parameter for logistic regressioin model in predicting heat stress using behavioral dataset
* Shahinfar 2019 – It a study use a series of predictors, such as the birth weight, age at slaughter, breeding values of fat composition of its ancesters caracse traits (thcarcases weight, intramuscular fat, etc). Considering the large size of feature space, it also deployed a nested crossvalidation where to selectoptimium number of decision trees and bootstrap sample fro the tree derivation. to avoid overfitting .

Block CV:

St-Pierre shared an similar observation when comparing dataset collected from different studies, which occurred under distinct time or envirnoemtn. The distinctness is reflected on data’s variance or value scale, making comparing them in literature difficult. The author defined such effect as “study effect”, and should be considered as random variable as it’s interested to calibrate the larger set of study effects including the unseen ones instead of the current observable effects. And the workaround is to use mixed model to calibrate the effect before making any inference of the dataset. In the prediction modeling context, since the feature matrix is the only known information throughout the prediction pipeline, the overestimation from the random CV could be alleviated by calibrating the feature matrix using the same mixed model idea before running the evaluating work.

De Oliverira 2020: Predict how the maize hybrids yield performa from different years, they conduct two CV system: CV1 and CV2, with the CV1 randomly assign hybridns and CV2 split by years, and model do show large different year do show more than 0.4 difference in r from year to year.

Another strategy is to incorporate environment covariates). Considering seasons are random effects that can not be observed in the history (since there is rarely two identifcal season in terms of in history in terms of its response to yield), instead of modeling the season effect directly, proxy variables that represent the season were obtained. In the study Cruz 2023, it breaksdown the evnrionemtn changes into 183 environment covariates into the models. That include cumulated thermal time to crop, soil water evaportation, leaf index, daily infiltration. By estimating the effect of these environment covariates, this should overcome the missing information from the unseen seasonal variation and reduce the performance drop from Random CV to Blcok CV.

Data simulation:

* show example study in agriculture using spectral data
  + Zhang 2019, - weed from rice
  + Su 2020 – weed detection
* Autocorrelation structure: 1100-2650nm, 235 bands. By Sun.

Metric:

* A case where CCC is outperform than r2
* Sensitivity vs specificty
  + Buczinski 2018 – Bovine respoiratory disease based on biometric indicator such as rectal temperature, eye discharge, etc
  + Lu 2017 – Wheat disease, strip rust and balck chaff from an image
  + He. 2021 - pig body weight group classification base on the feeding data
* Example use RMSE and MAE for outliers
* F1
  + Haque 2023 maize leaf blight detection, the evaluated dataset is imbalance between the healthy leave and the disease leaves in a ratio of 1 to 2, with an accuracy of 99.02%, it also report the F1-score of 97.49% to ensrue that the accuracy is not inflated due ot the imbalanced.
* F2
  + Minni (2024) Breast cancer diagnosis, no actual cancer patient should be left unpredicted (high TPR). Hcne using f2 to select the best machine learning strategy
  + Prasetiyo 2024: Same strategy go to fraud detection where missing a positive case is costly
* MCC
  + Others focus on positive predictions
  + MCC is the only metric that focus on both samples and less sensitive to change balnce.
  + Not sensitive to the change ot blance dataset but also consider multiple sample distribution.
  + Becker 2021- Predicting dairy cattle heat stress using the scoring system:temperature-humidity index, rsspiration rate, and other behavioral indicator such as lycing time.
    - It has multiple dataset with different balance levels from mild imbalce (119to 153) to several imbalance (265 non – 50 heat). It also pointed out that reporting F1 mostly reflect the classi distribution, not the goddness ofth e classifier.