Comparing Explanations between Random Forests and Artificial Neural Networks

Lee Harris and Marek Grzes

The University of Kent,
Canterbury,
United Kingdom

Explanations

Explanations

- How does the model make a decision for a particular input?

Explanations

- How does the model make a decision for a particular input?
  - Displayed through importance scores, rules, heatmaps, ect.

Explanations

- How does the model make a decision for a particular input?
  - Displayed through importance scores, rules, heatmaps, ect.

- Local Fidelity

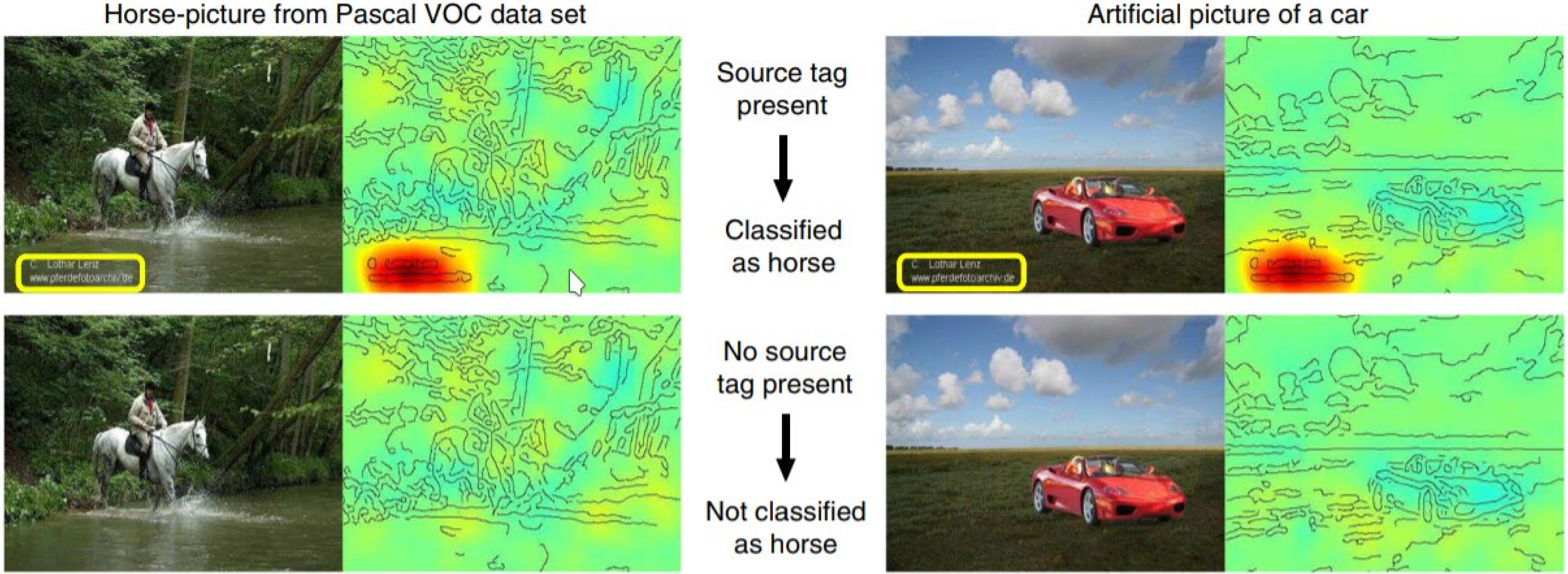
Explanations

- How does the model make a decision for a particular input?
  - Displayed through importance scores, rules, heatmaps, ect.

- Local Fidelity

- Why do we want to know this?

Explanations: Incorrect Behaviour

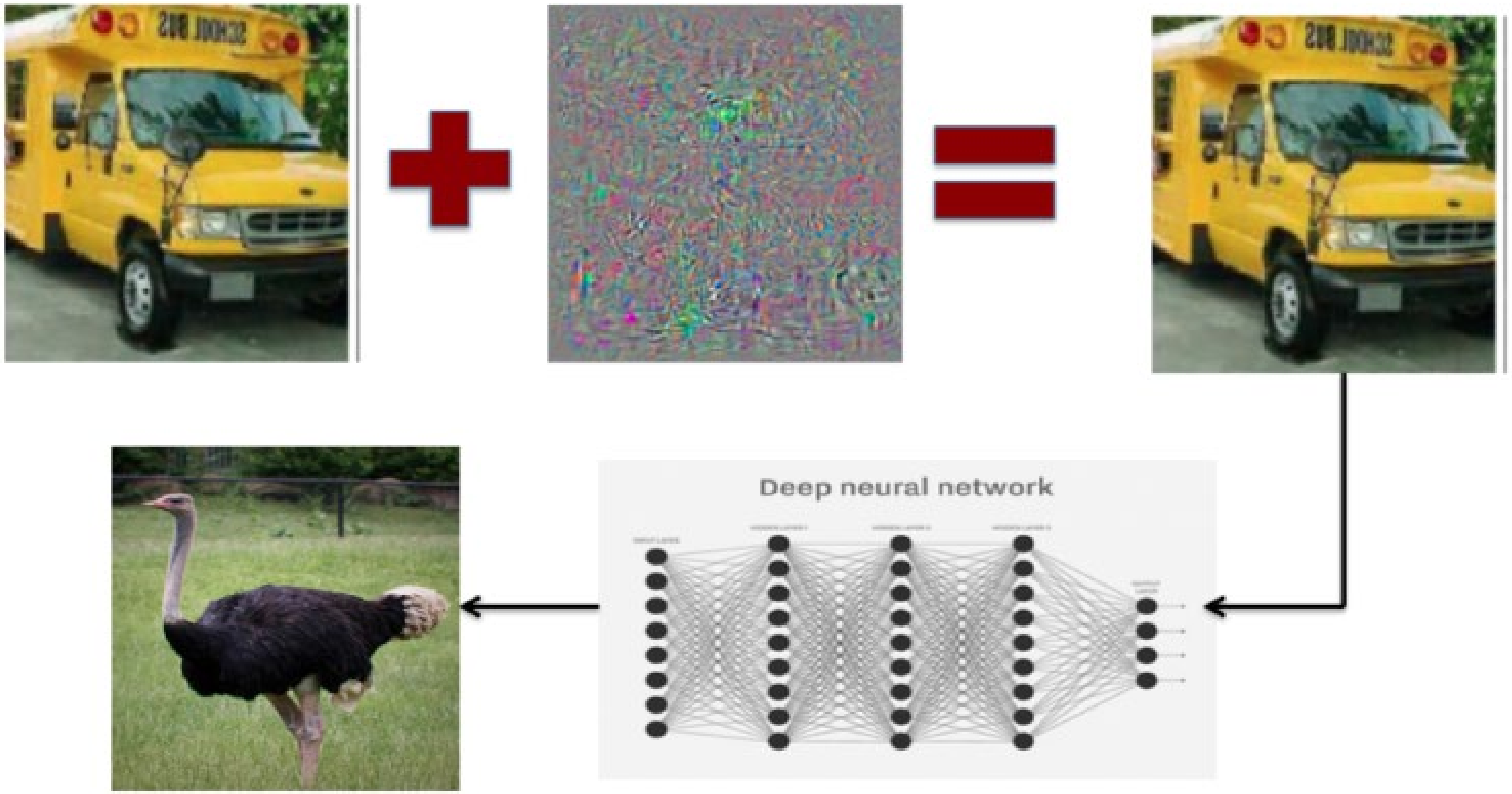


Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), 1096.

Explanations: Cheating in Video Games

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), 1096.

Explanations: Adversarial Attacks



Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R., 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Explanations:

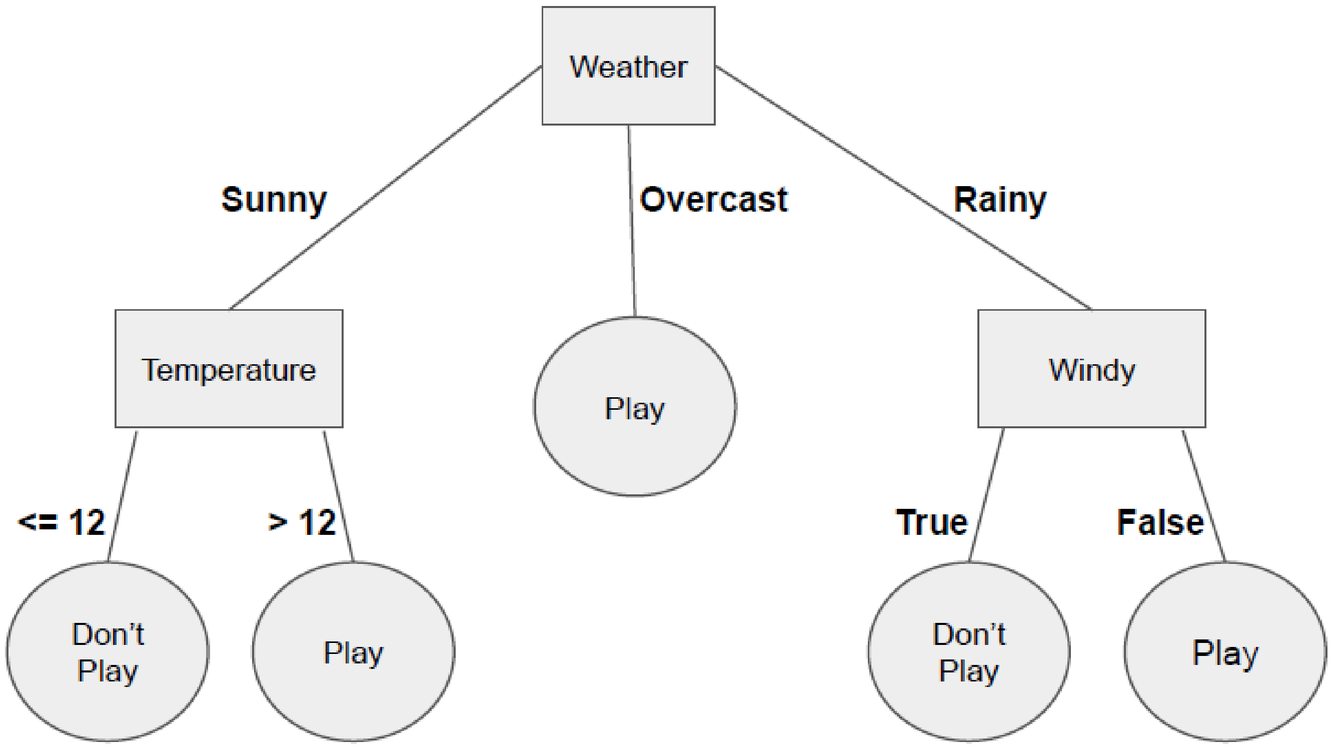
- US predictive policing: Rubin, J., 2010. Stopping crime before it starts. *Los Angeles Times*, *21*.

- Self driving cars and racist machines: Wilson, B., Hoffman, J. and Morgenstern, J., 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*.

- Admissions to Berkeley College: Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.P., Humbert, M., Juels, A. and Lin, H., 2017, April. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 401-416). IEEE.

- Interesting patterns

Decision Trees



Rules:

1) If(Weather = Sunny And Temp. <= 12) Then Don't Play
2) If(Weather = Sunny And Temp. > 12) Then Play
3) If(Weather = Overcast) Then Play
4) If(Weather = Rainy And Windy = True) Then Don't Play
5) If(Weather = Rainy And Windy = False) Then Play

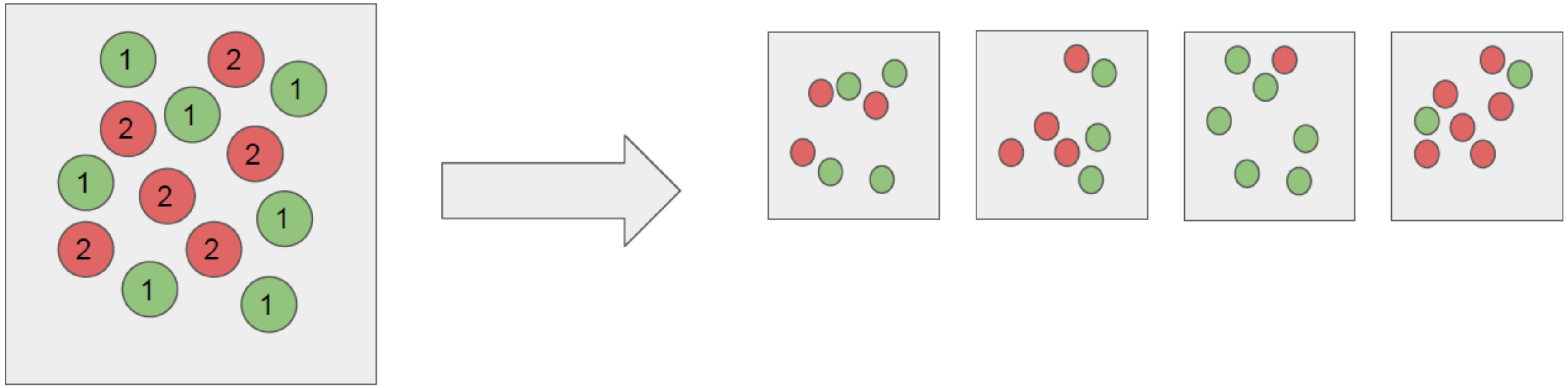
- Can be represented as a series of rules
- Not the best predictors

University of Kent

Random Forests

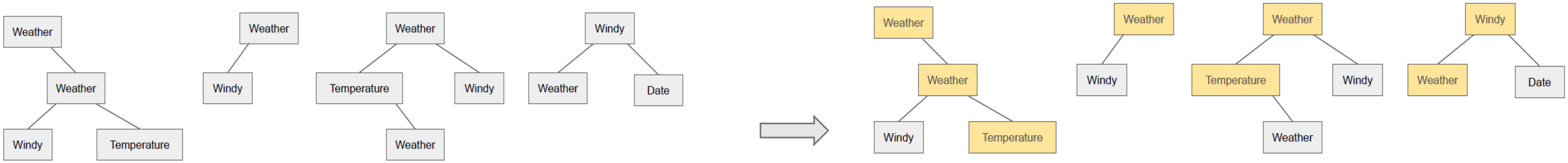
- Ensemble of many trees

- Grey-box



Intervention in Prediction Measure (IPM)[3]

- Each feature importance is the average feature-frequency across every traversed path, averaged over the entire ensemble of trees

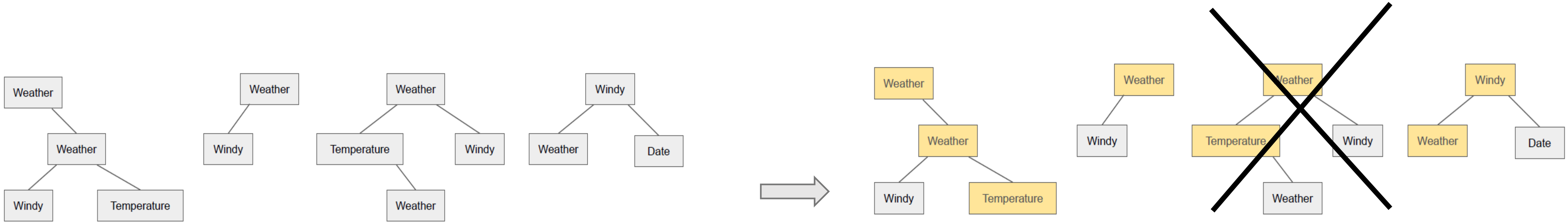


Frequencies
Weather = 6
Windy = 4
Temperature = 2
Date = 1

Frequencies
Weather = 5
Temperature = 2
Windy = 1
Date = 0

Adjusted Intervention in Prediction Measure

- Our adaption over just trees in the majority

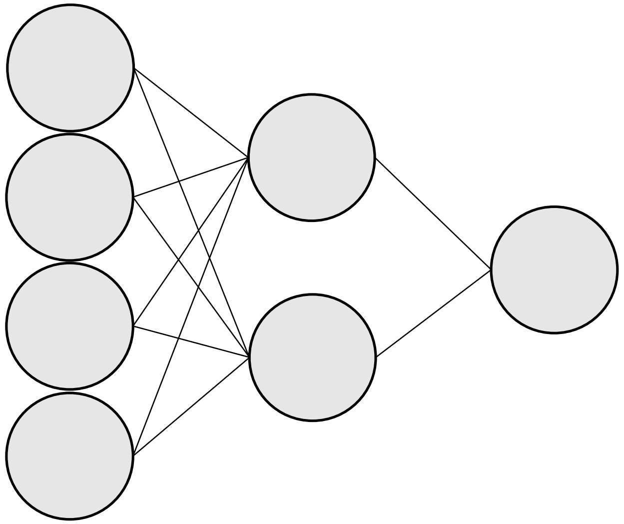


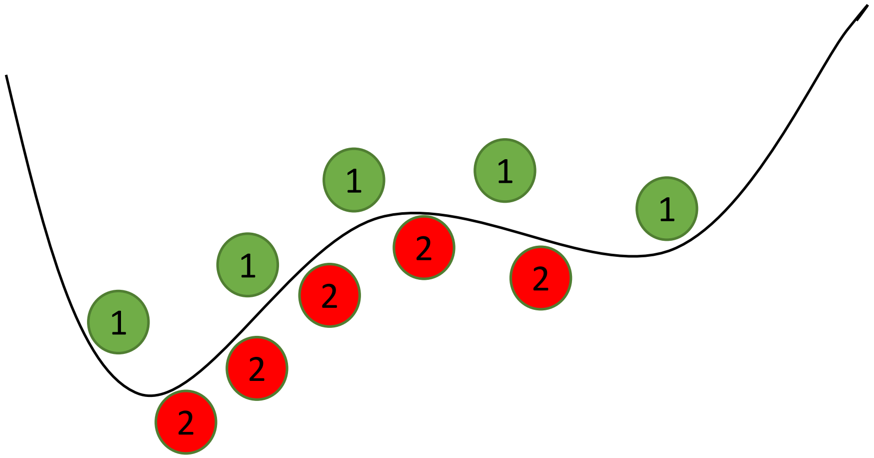
Frequencies
Weather = 6
Windy = 4
Temperature = 2
Date = 1

Frequencies
Weather = 4
Windy = 1
Temperature = 1
Date = 0

Artificial Neural Network

- High predictive performance…

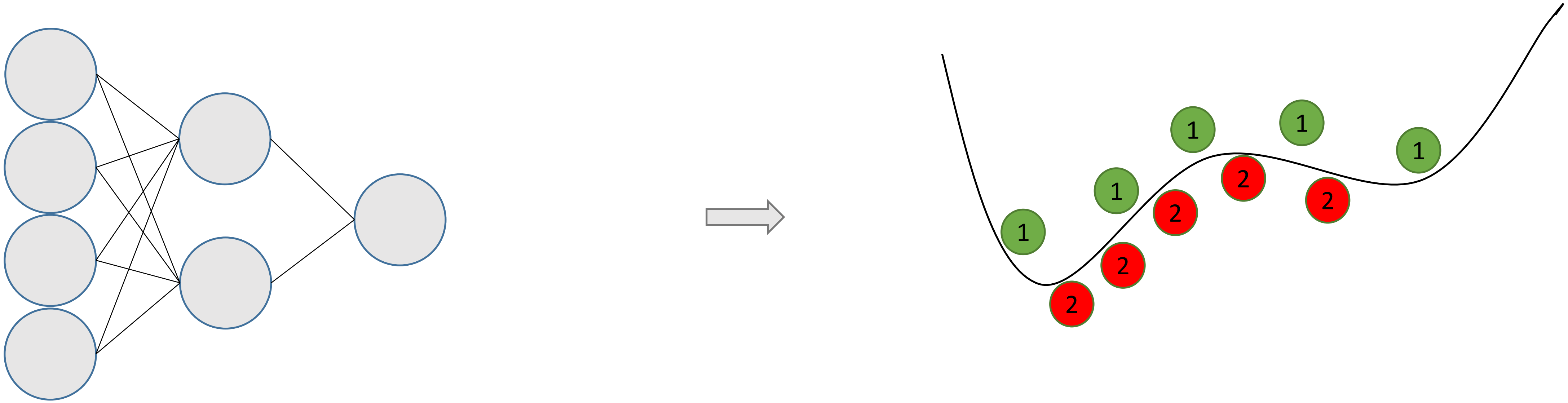




Artificial Neural Network

- High predictive performance…

… but complex reasoning



Sensitivity Analysis

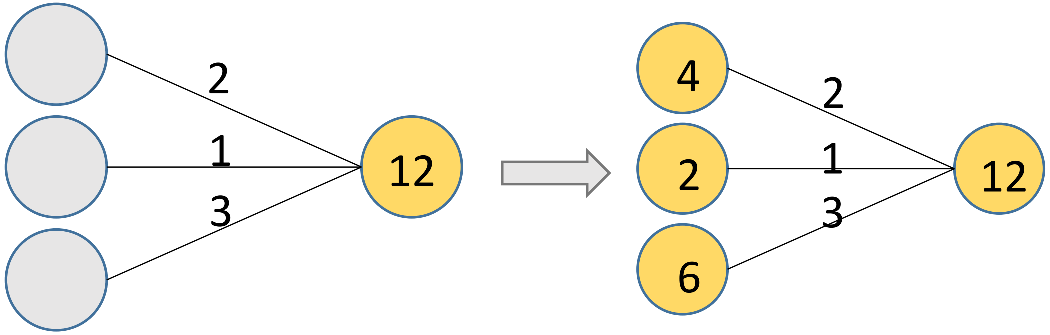
- Natural

- Established

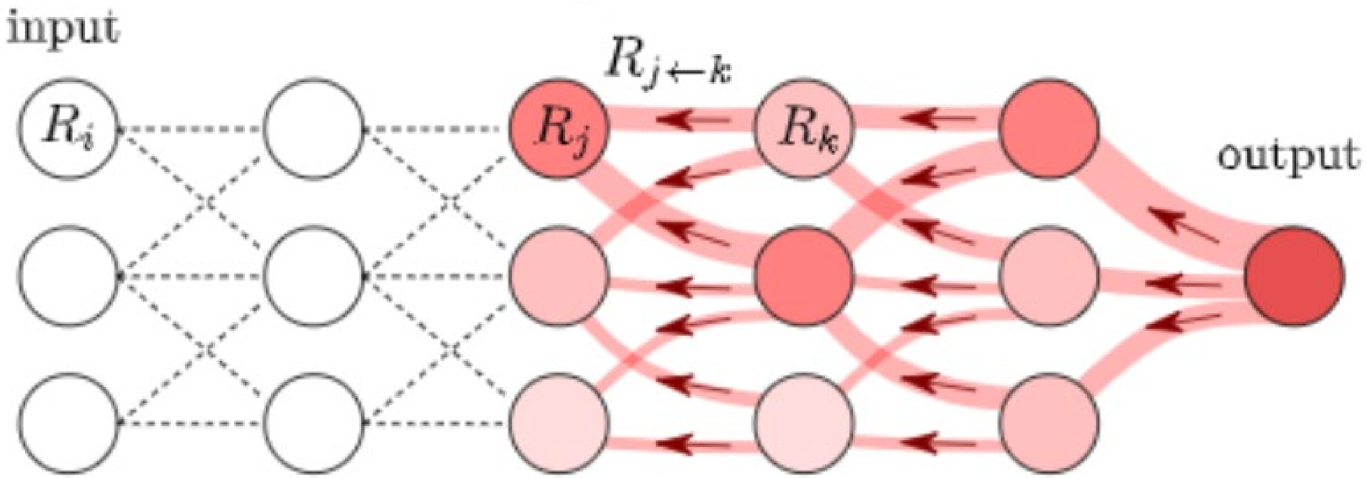
$$S_i(X) = \sqrt{\sum_{k=1}^{|o|} \left(\frac{\partial o_k}{\partial X_i}\right)^2} = \left\|\frac{\partial o}{\partial X_i}\right\|_2$$

Layerwise Relevance Propagation[4]

- Backpropagate activation



$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$



Montavon, G., Samek, W. and Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, pp.1-15.

Motivations

Motivations

- Decision trees are transparent

Motivations

- Decision trees are transparent
- Therefore, random forests must have some transparency

Motivations

- Decision trees are transparent
- Therefore, random forests must have some transparency
- If the explanations extracted from artificial neural networks correlate, these must also have some level of transparency

Motivations

- Decision trees are transparent
- Therefore, random forests must have some transparency
- If the explanations extracted from artificial neural networks correlate, these must also have some level of transparency
- Both models are nonlinear, but have significantly different structure and decision making.

Our Work

- The first comparison of explanations between random forests and artificial neural networks

- High-level features

- Real and Synthetic datasets

Base Method

- Randomly sample a unique instance
- Generate 3 different models on the rest of the data
- Explain these models with each explanation method
- Discretise each explanation (Most importance feature = rank 1)
- Repeat this t (100) times
- Plot the average feature rank

```
Research(int trials){
    feature_ranks = [trials]
    for(t in trials){
        instance = sample(instances)
        unbalanced_forest = new RF(instances – instance)
        rf = unbalanced_forest.explain(instance)
        balanced_forest = new CF(instances – instance)
        cf = balanced_forest.explain(instance)
        neural_network = new ANN(instances – instance)
        network = neural_network.explain(instance)
        feature_ranks[t] = [rf, cf, network]
    }
average_rank = trials / sum(feature_ranks, column)
plot(average_rank)
}
```

Synthetic Data

Scenario	Instances	Features
S1	300	6
S2	3000	12
S3	1500	30

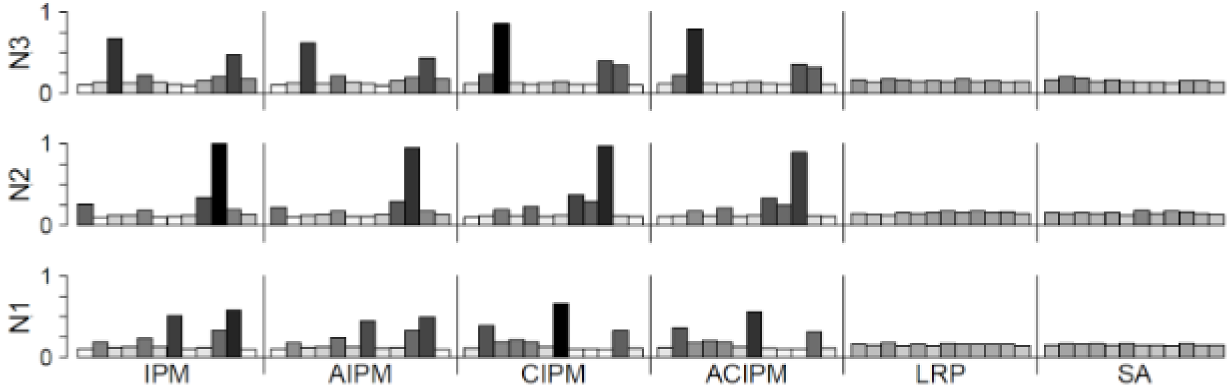
- 3 different dataset dimensions (i.e. ‘scenario’)

- Each of these explores 4 problems
    1. No important features
    2. A single important feature
    3. Two important features
    4. Relative feature importance

- 3 different noise levels: 10%; 20%; 30%

Results I - Simulated

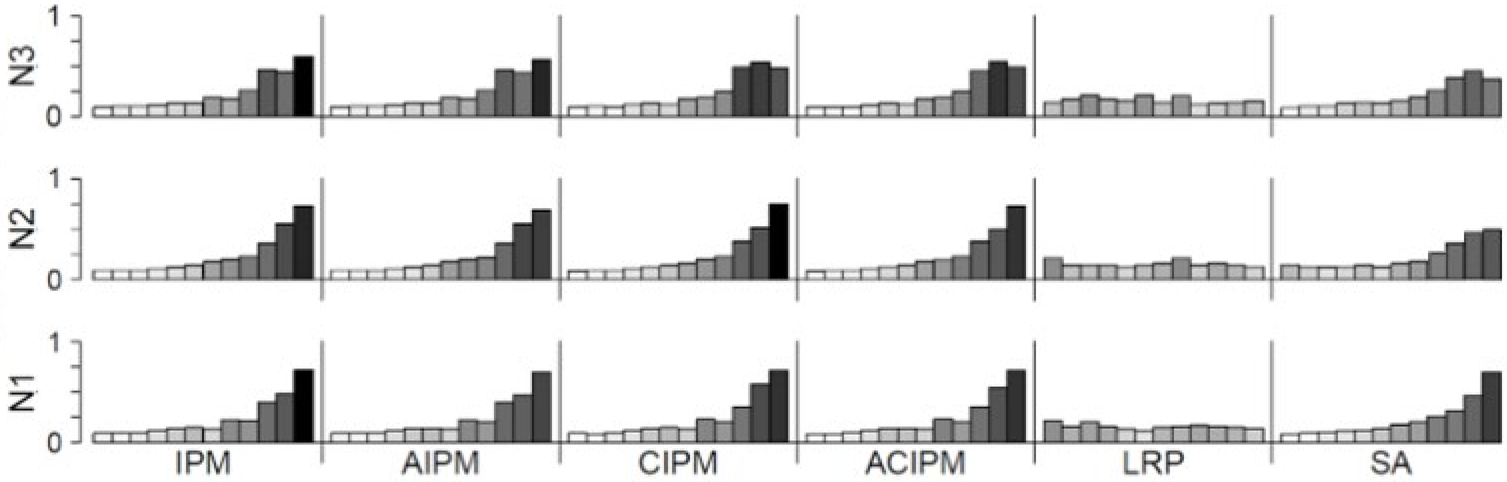
No Important Features (Baseline)



Scenario 2:

- 3000 Instances
- 12 Features
- 3 Noise levels

Relative Feature Importance



University of Kent

Correlations

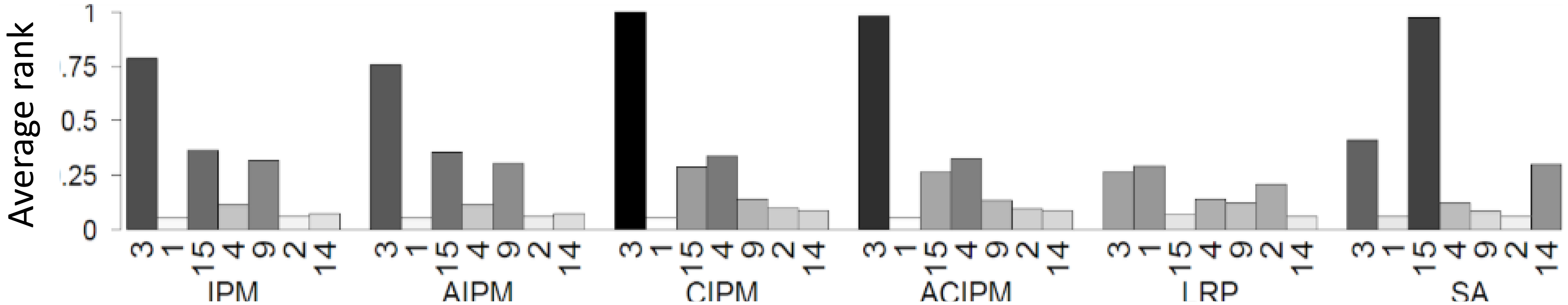
- If each feature in each explanation is ranked, it is possible to compare them

- This is only reported for the real data

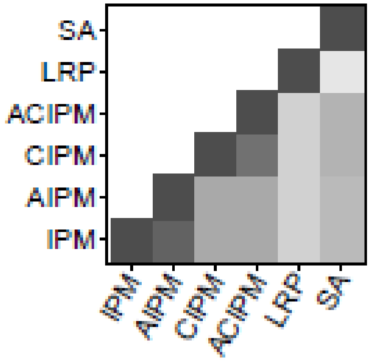
- These were nearly all positive

Results II - Real Data

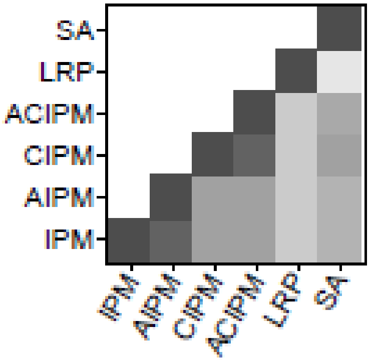
20 Continuous Features



Diabetic Retinopathy[1]



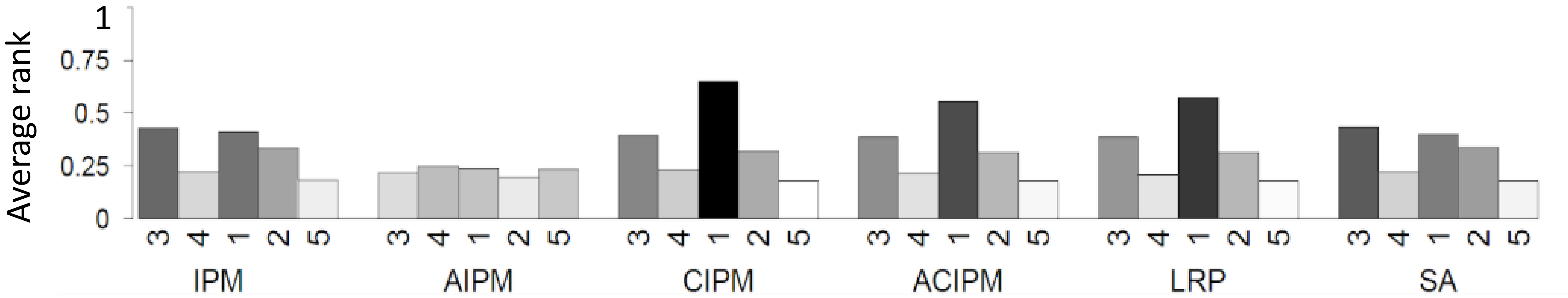
(a) Kendall's τ



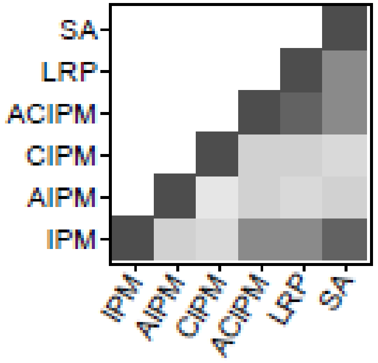
(b) Spearman's ρ

Results II - Real Data

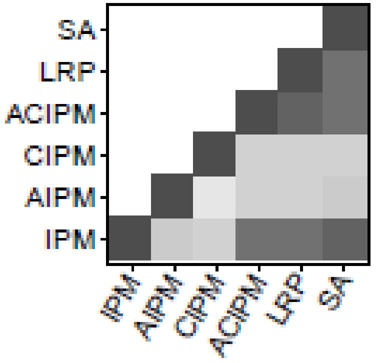
10 Discrete Features



Website Phishing[2]

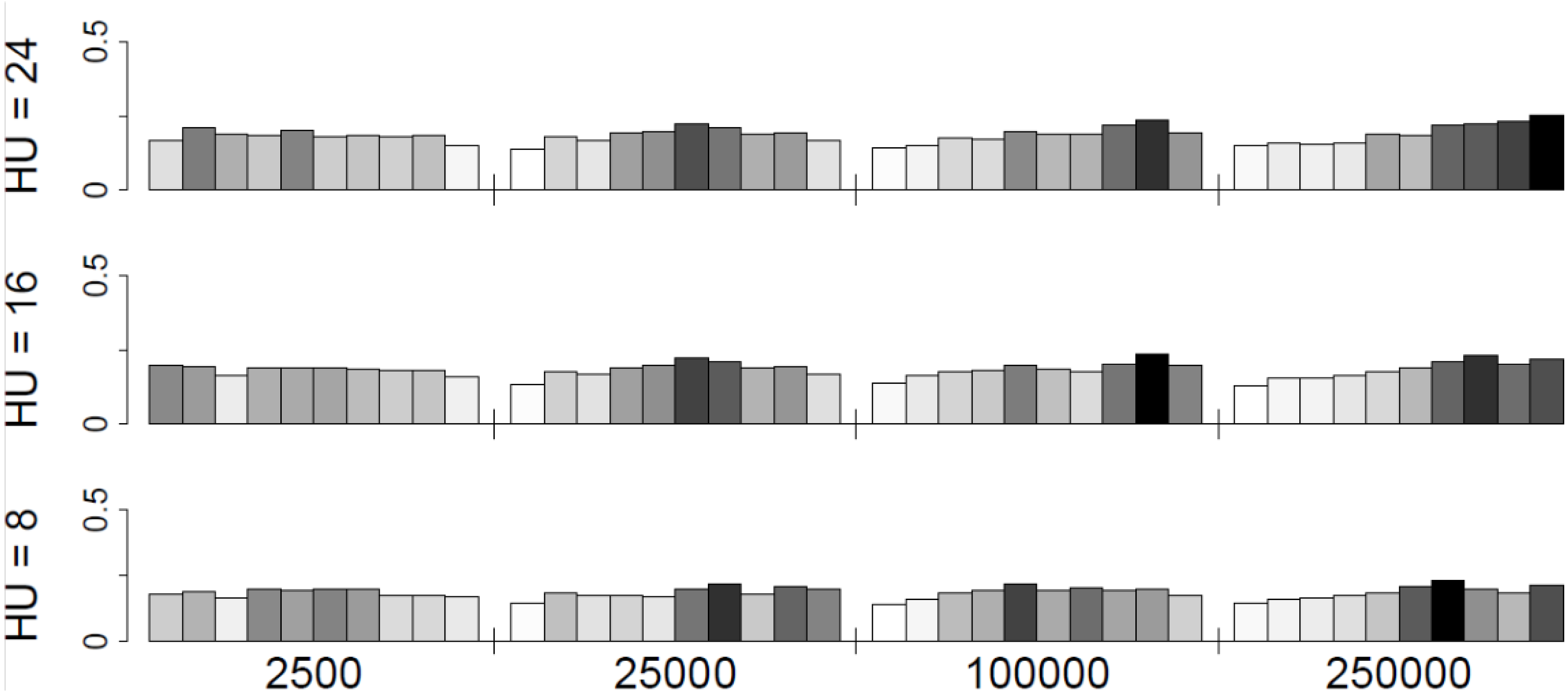


(a) Kendall's τ



(b) Spearman's ρ

Results III - Further Testing



Conclusion

- Datasets with fewer features correlate more

- High predictive accuracy does not guarantee similar explanations

- Explores the IPM method applied to higher dimensionality

- The certainty assigned by Layerwise Relevance Propagation increases with the number of hidden units

- Balanced Random Forests appear more promising for explainability

Future Work

- Exploration of other random forest and network architectures
    - Deeper networks
    - Other RF variants

- Additional datasets
    - Low level features
    - Extracting features

References

[1] - Diabetic Retinopathy: Antal, B. and Hajdu, A., 2014. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-based systems*, *60*, pp.20-27.

[2] - Website Phishing: Abdelhamid, N., Ayesh, A. and Thabtah, F., 2014. Phishing detection based associative classification data mining. *Expert Systems with Applications*, *41*(13), pp.5948-5959.

[3] - Intervention in Prediction Measure: Epifanio, I., 2017. Intervention in prediction measure: a new approach to assessing variable importance for random forests. BMC bioinformatics, 18(1), p.230.

[4] - Layerwise Relevance Propagation: Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R. and Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), p.e0130140.

[5] - Conditional Inference Forest: Hothorn, T., Hornik, K. and Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), pp.651-674.

[6] - Quinlan, J.R., 2014. C4. 5: programs for machine learning. Elsevier.

Thank you For Listening

Lee Harris - https://www.cs.kent.ac.uk/people/rpg/lh558/

Marek Grzes - https://www.cs.kent.ac.uk/people/staff/mg483/

The University of Kent - https://www.kent.ac.uk/