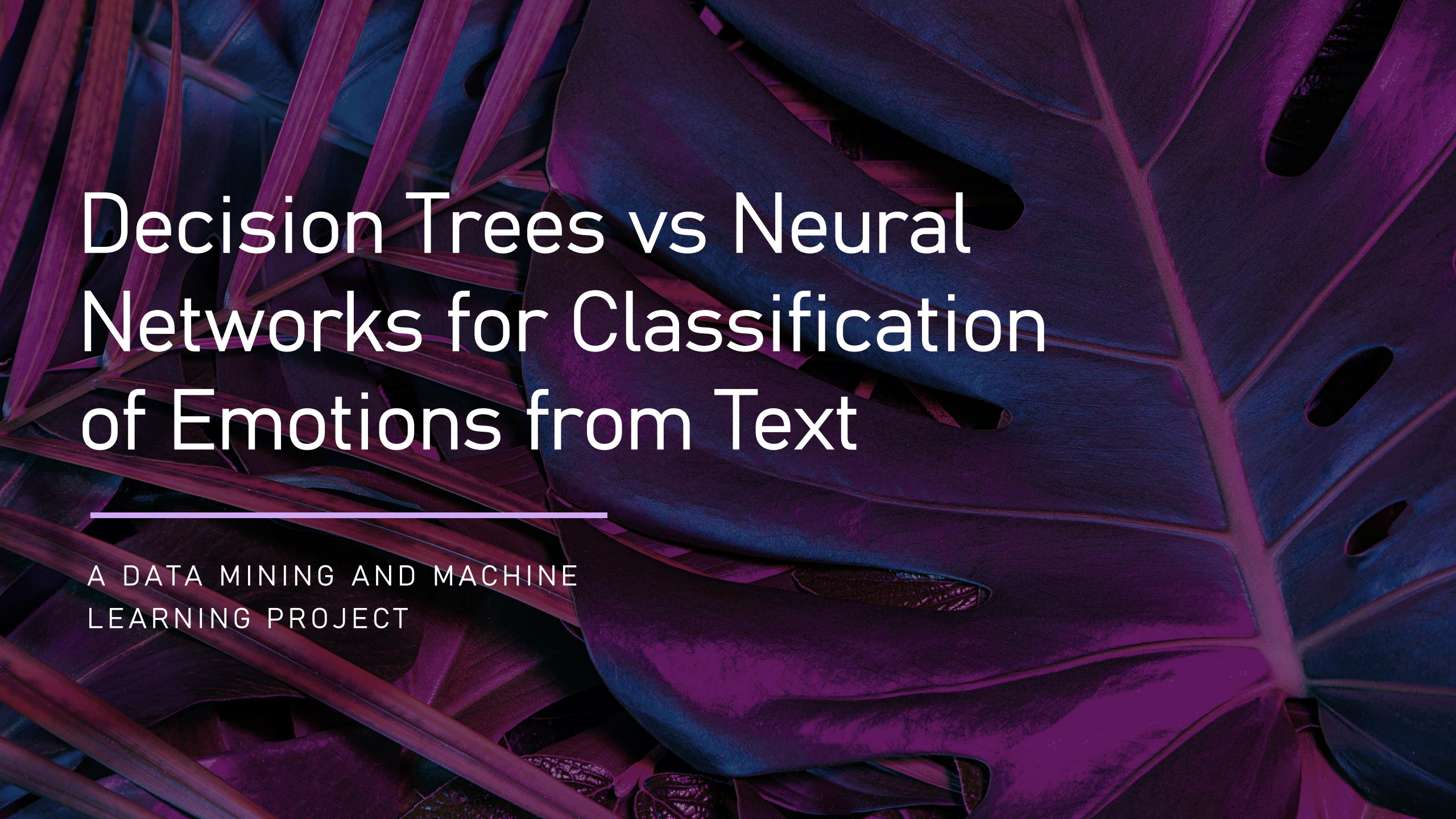
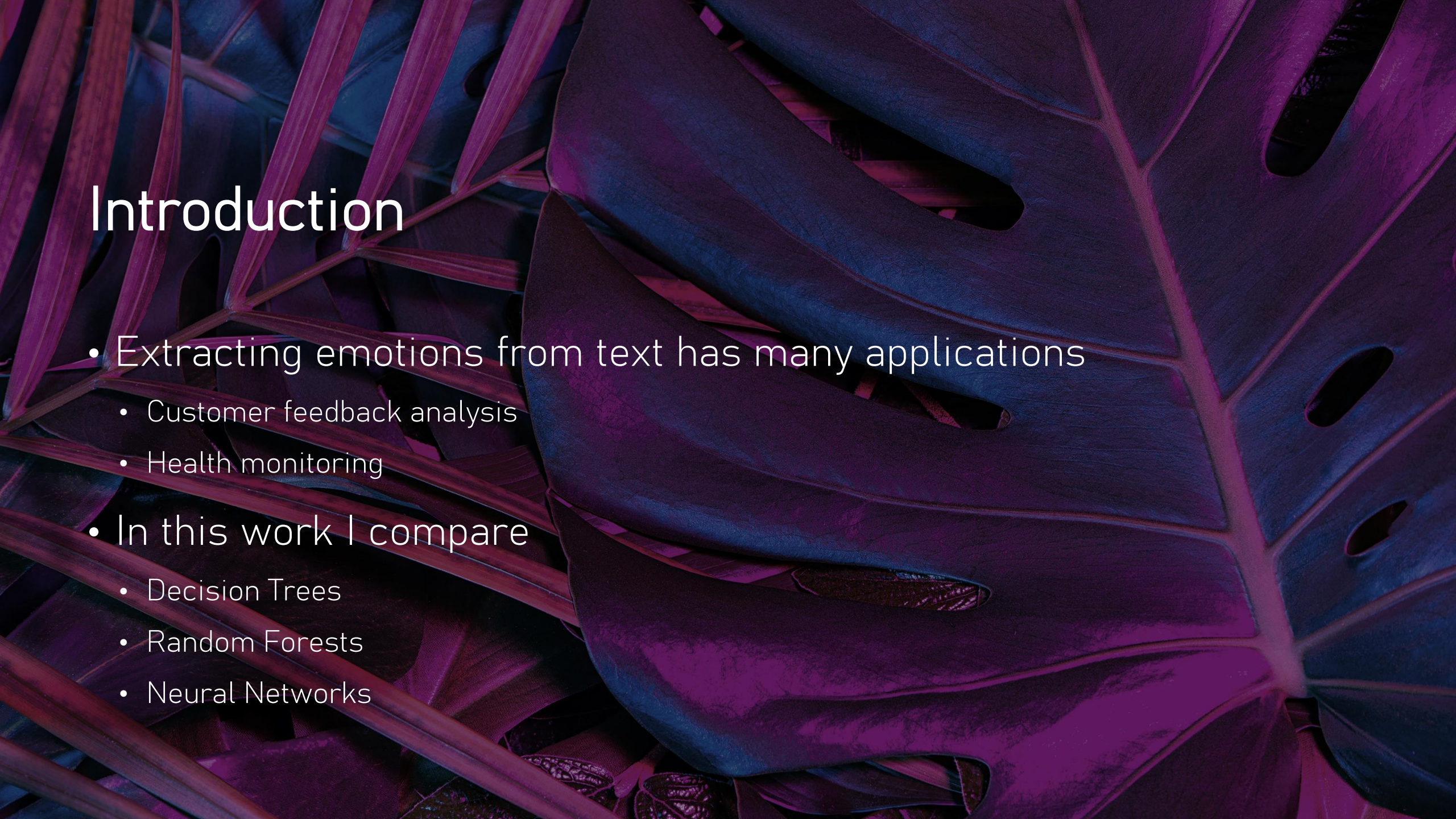


Decision Trees vs Neural Networks for Classification of Emotions from Text

A DATA MINING AND MACHINE LEARNING PROJECT

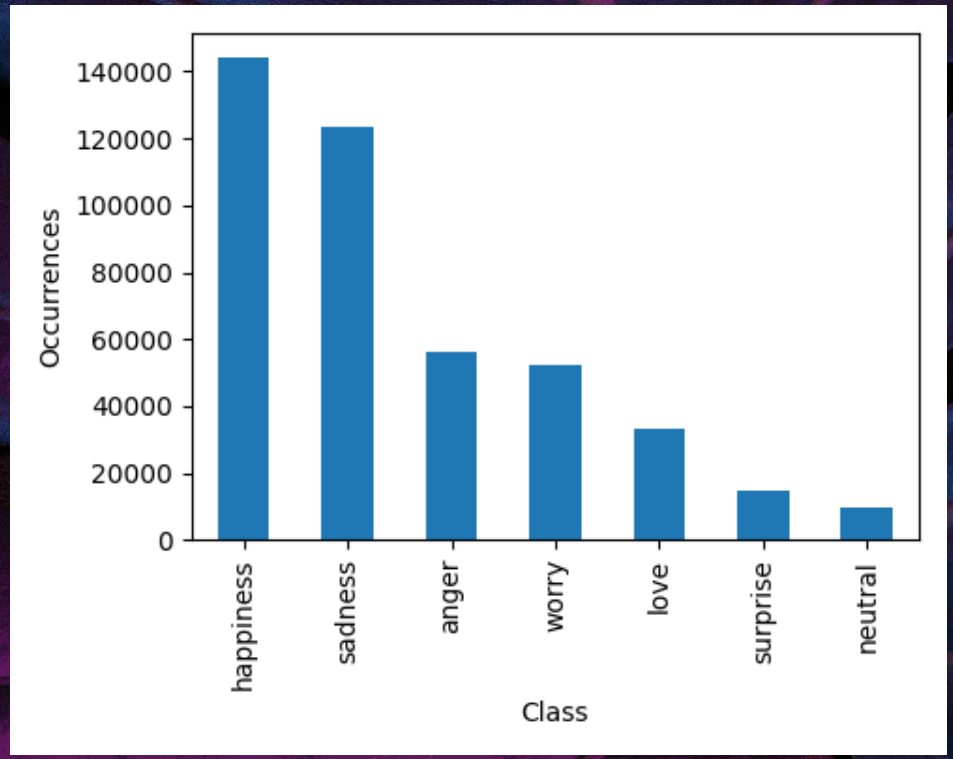


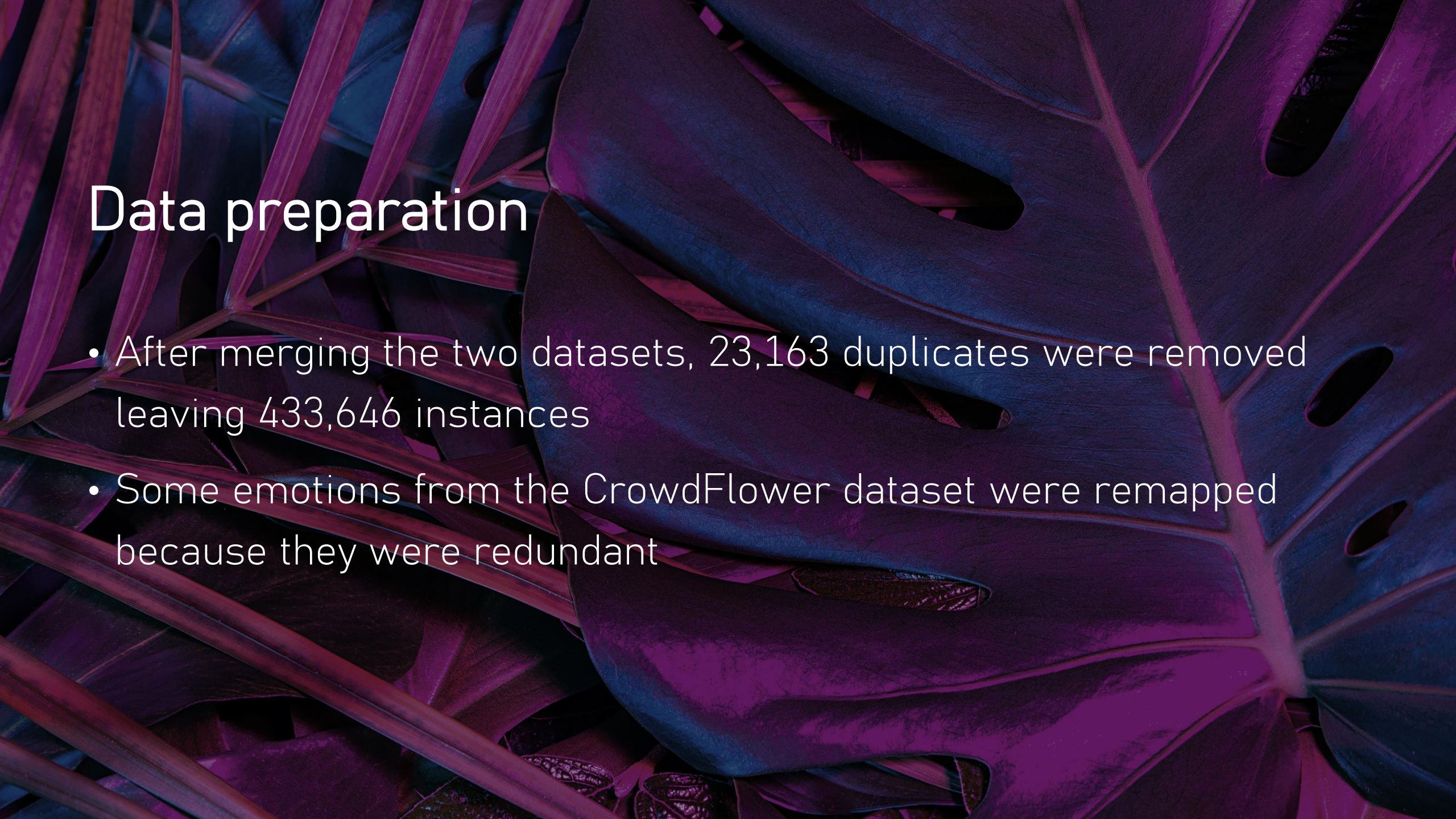
Introduction

- Extracting emotions from text has many applications
  - Customer feedback analysis
  - Health monitoring
- In this work I compare
  - Decision Trees
  - Random Forests
  - Neural Networks

Data

- Two datasets of labelled tweets were merged
  - CrowdFlower (40,000 instances)
  - Emotions (400,000 instances)
- The resulting dataset has 7 emotions and is unbalanced
- A weight for each class was used to train the models





Data preparation

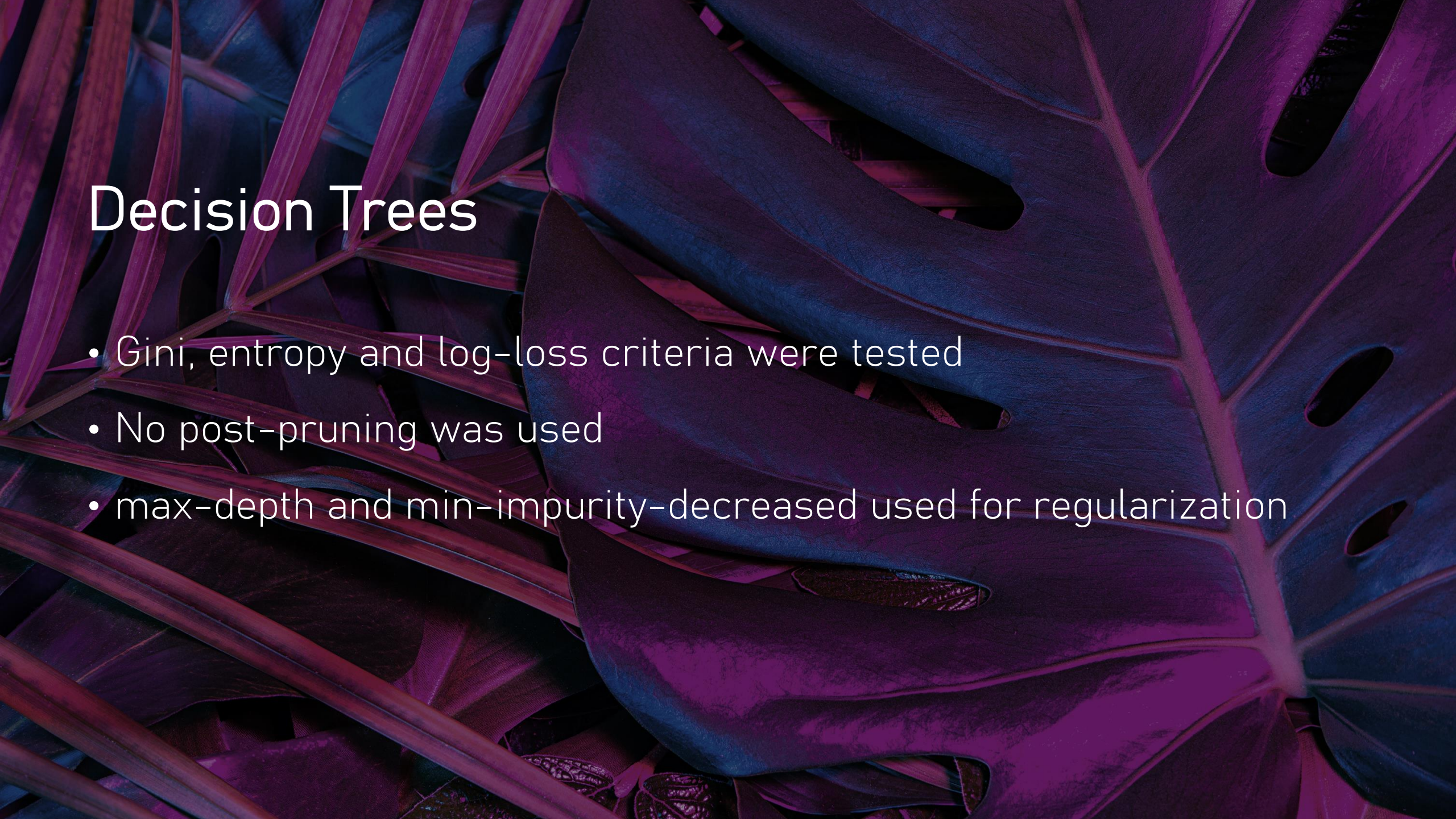
- After merging the two datasets, 23,163 duplicates were removed leaving 433,646 instances
- Some emotions from the CrowdFlower dataset were remapped because they were redundant

Preprocessing

- Different techniques were tested
  - BoW
    - Binary
    - TFIDF
  - Word Embedding
    - Learned
    - Pretrained GloVe

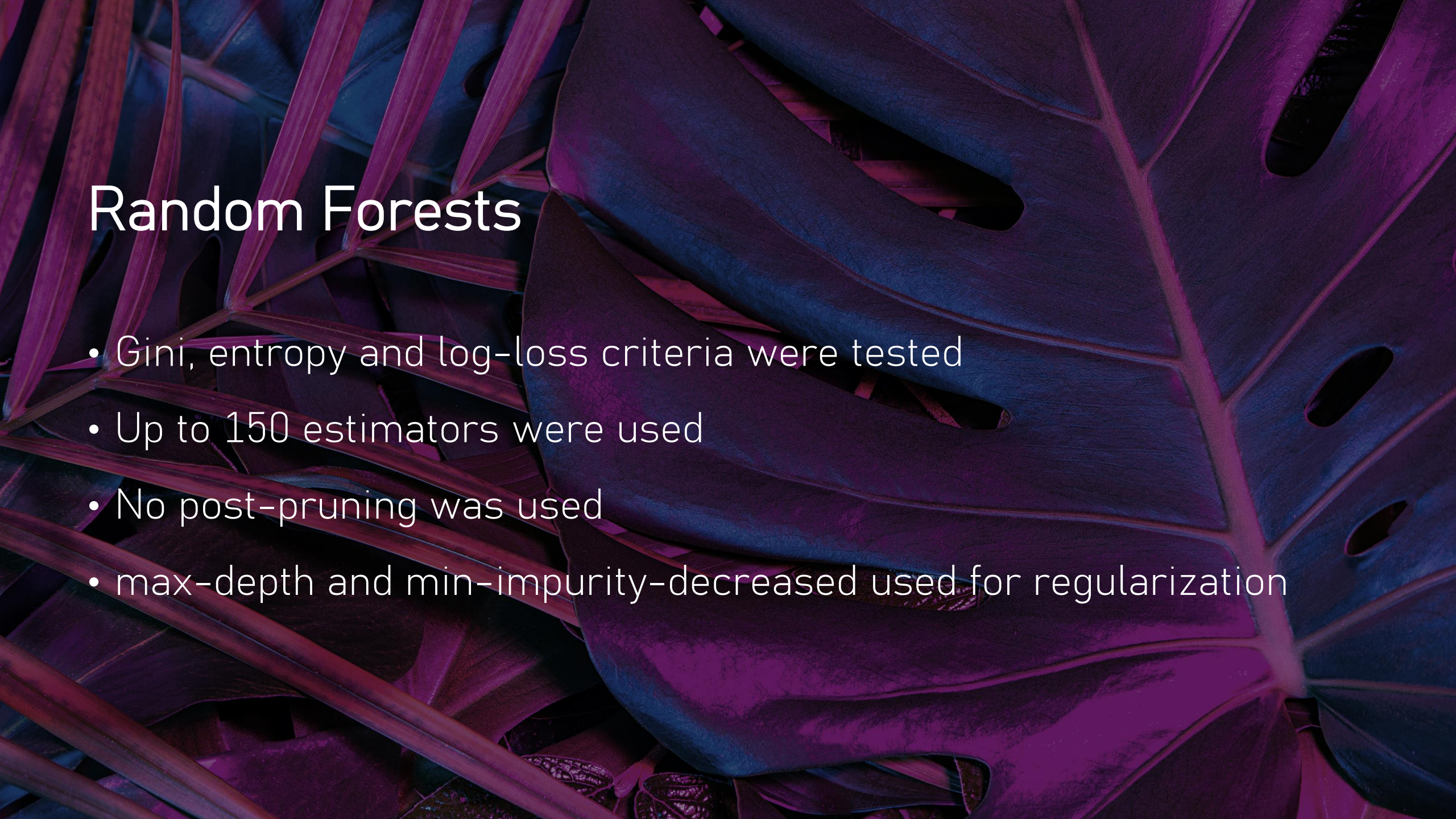
Models

- Three kinds of models were used
  - Decision Trees
  - Random Forests
  - Neural Networks



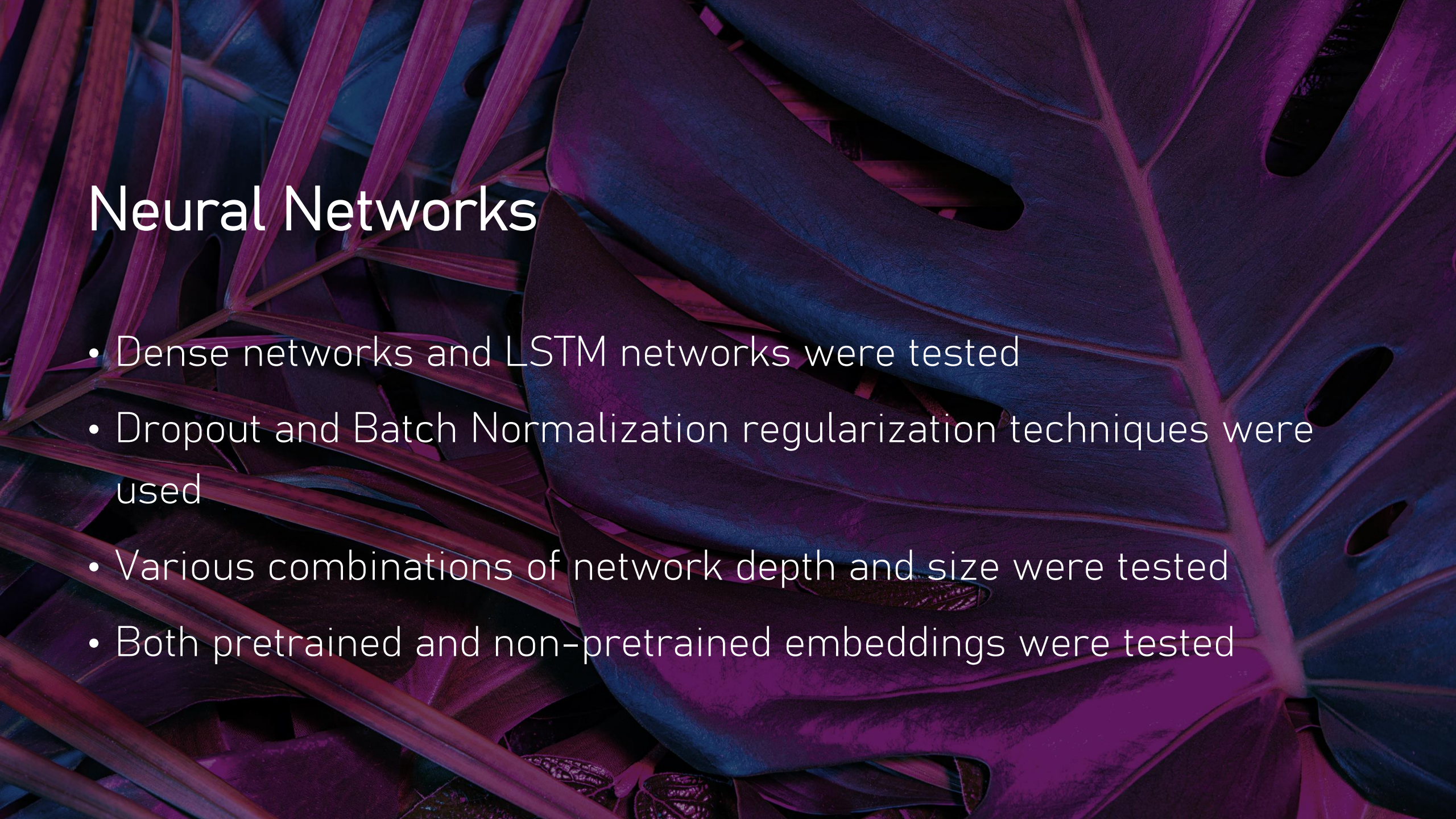
Decision Trees

- Gini, entropy and log-loss criteria were tested
- No post-pruning was used
- max-depth and min-impurity-decreased used for regularization



Random Forests

- Gini, entropy and log-loss criteria were tested
- Up to 150 estimators were used
- No post-pruning was used
- max-depth and min-impurity-decreased used for regularization



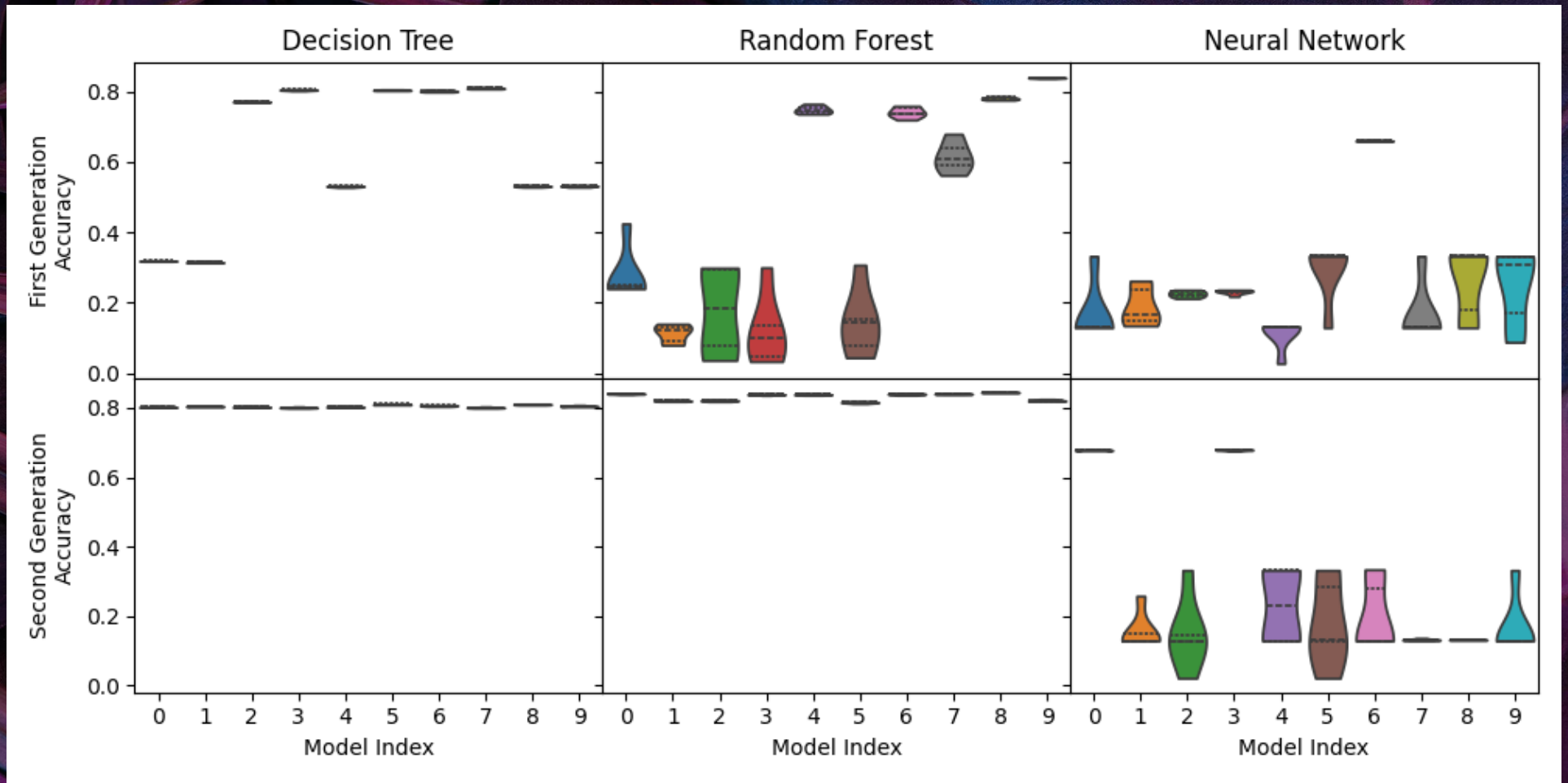
Neural Networks

- Dense networks and LSTM networks were tested
- Dropout and Batch Normalization regularization techniques were used
- Various combinations of network depth and size were tested
- Both pretrained and non-pretrained embeddings were tested

Model Selection

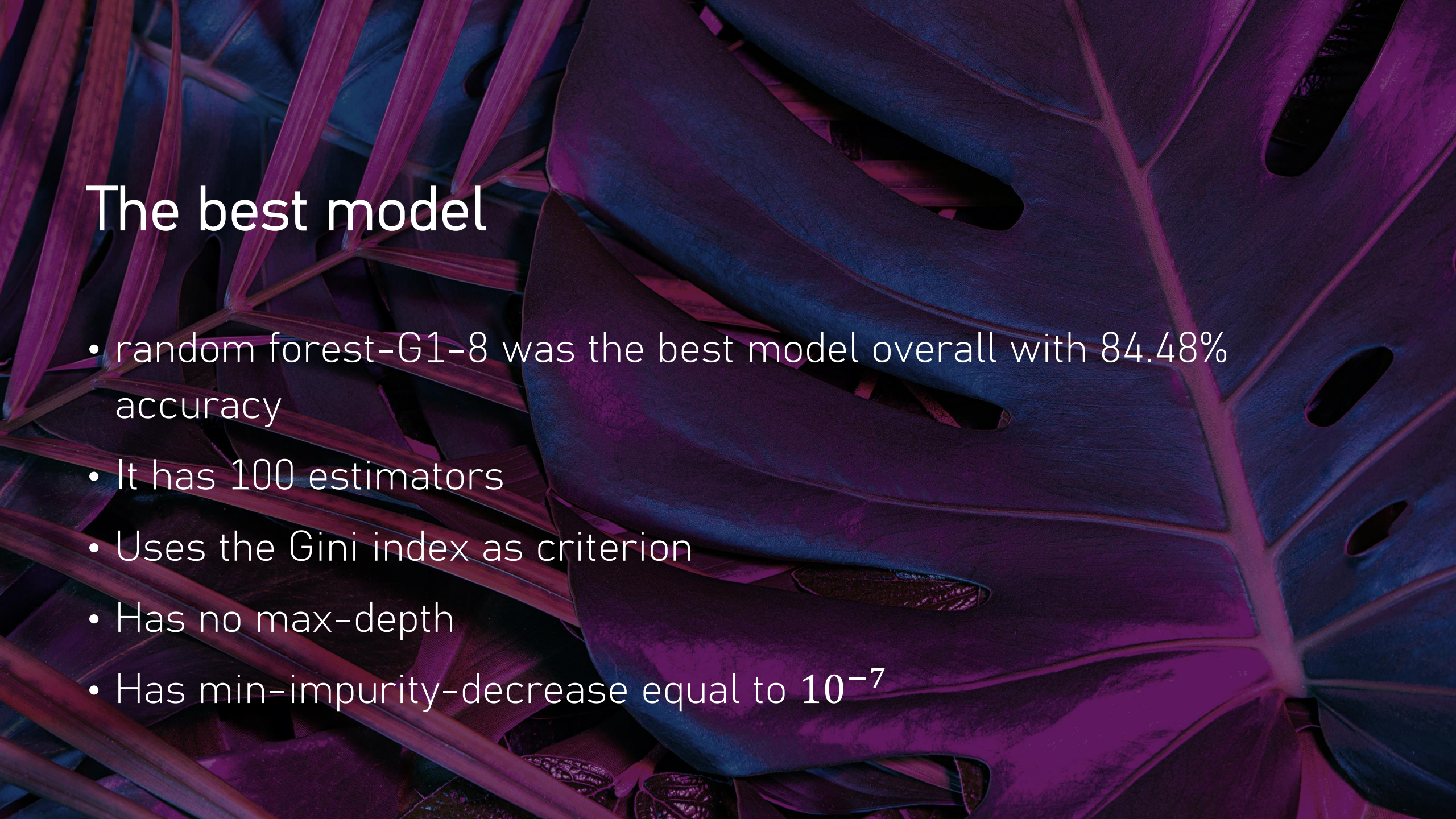
- Random search was used
- Two generations of 10 models for each kind were tested

Results


Decision Tree
Random Forest
Neural Network
First Generation Accuracy
Second Generation Accuracy
0.0
0.2
0.4
0.6
0.8
0 1 2 3 4 5 6 7 8 9
Model Index

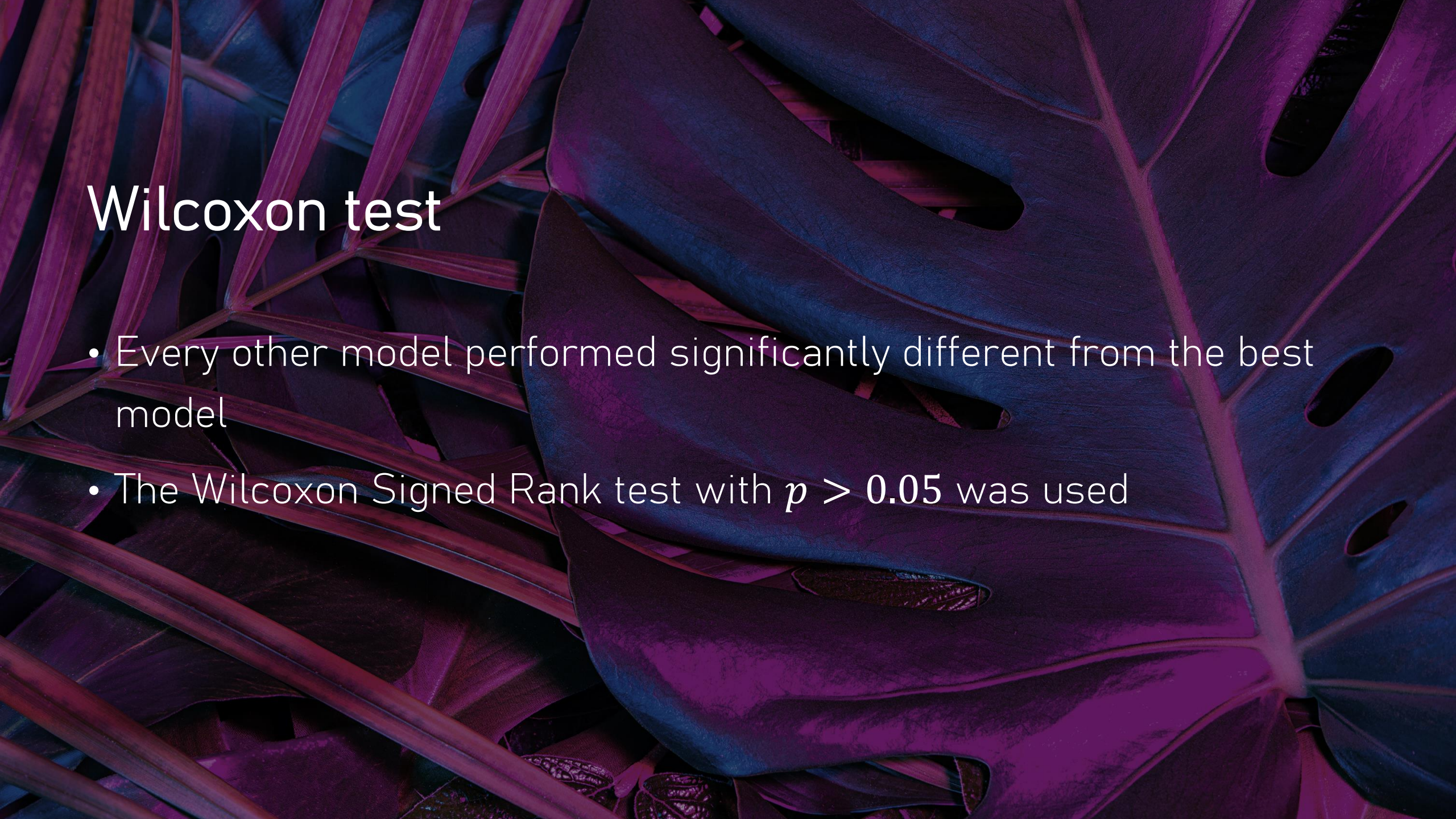
Best model for each kind

Model name	Mean Accuracy
decision tree-G1-5	81.02%
random forest-G1-8	84.48%
neural network-G1-3	67.87%



The best model

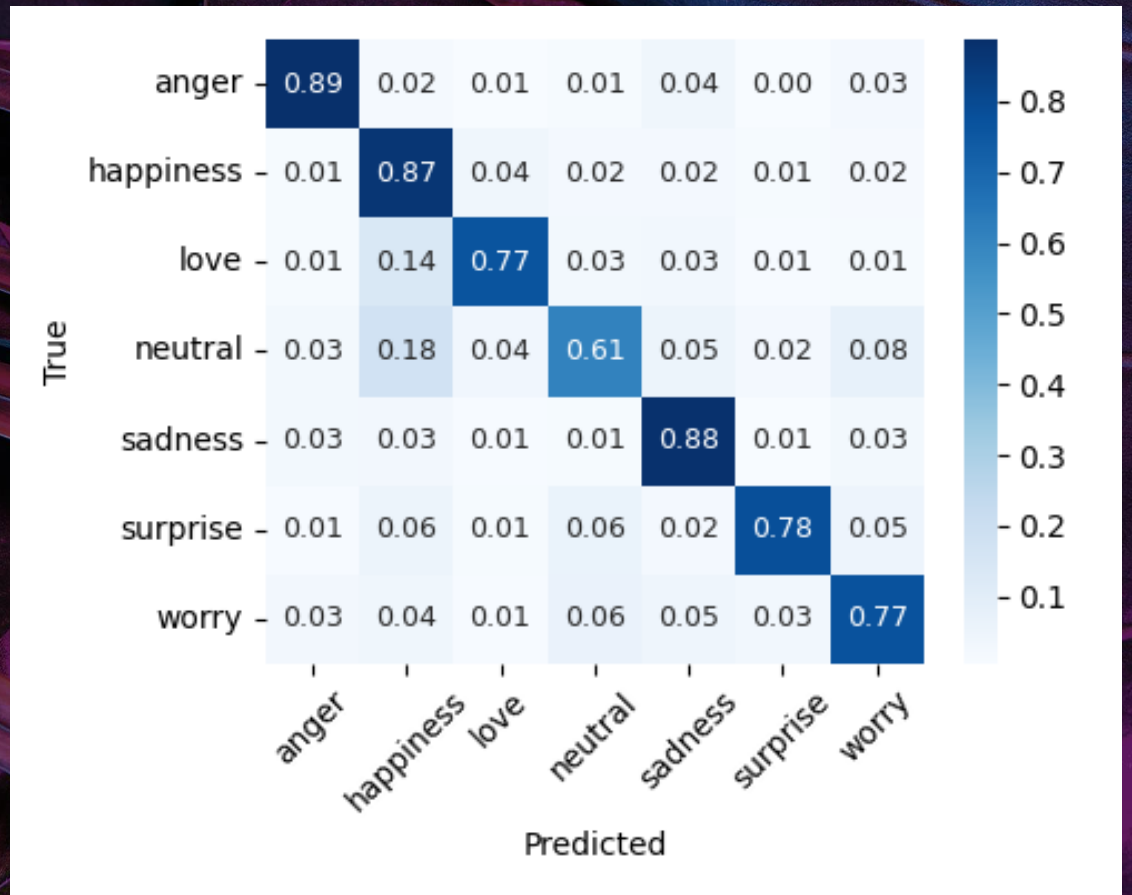
- random forest-G1-8 was the best model overall with 84.48% accuracy
- It has 100 estimators
- Uses the Gini index as criterion
- Has no max-depth
- Has min-impurity-decrease equal to 10^{-7}



Wilcoxon test

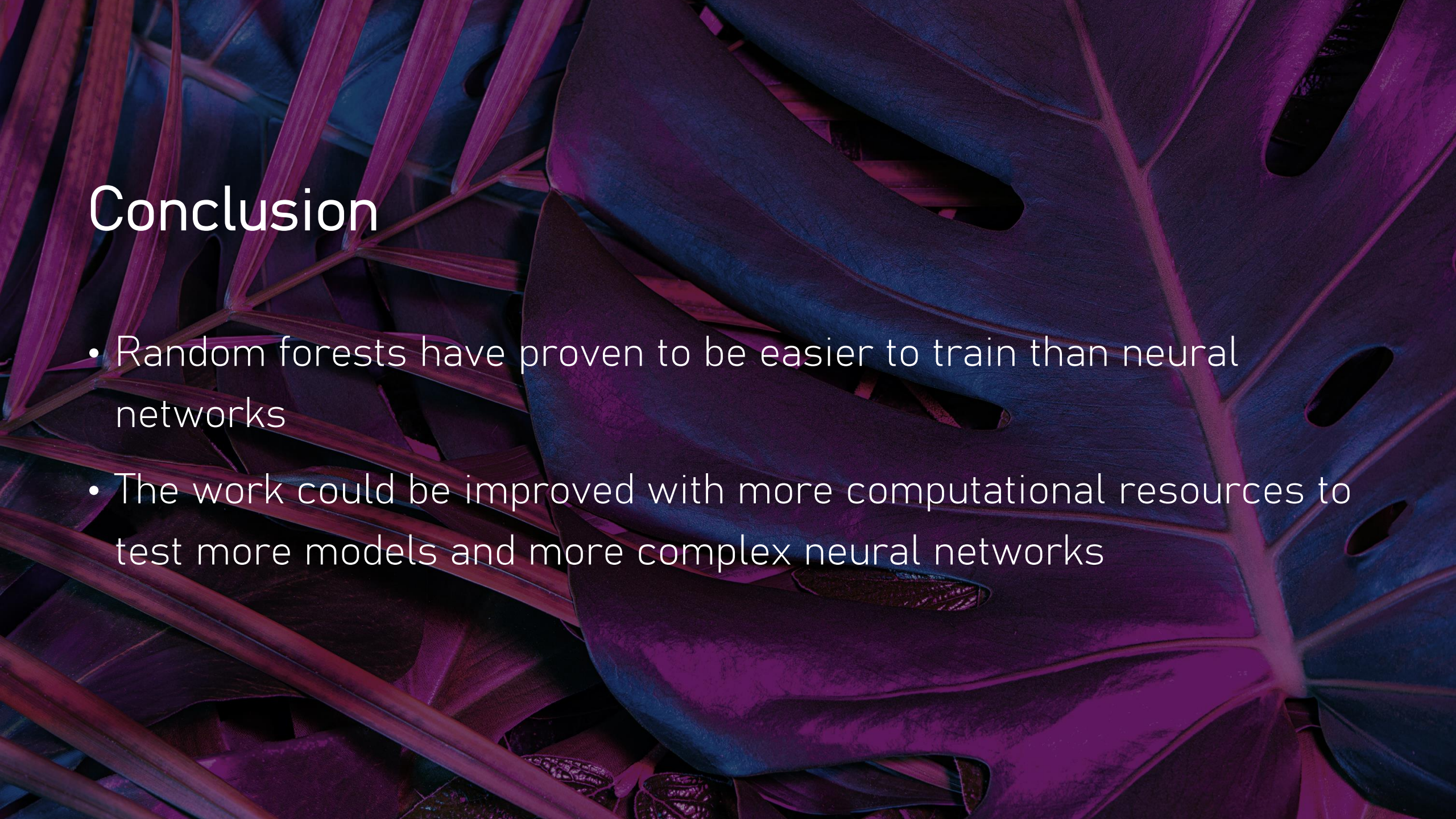
- Every other model performed significantly different from the best model
- The Wilcoxon Signed Rank test with $p > 0.05$ was used

random forest-G1-8 confusion matrix



Comparison with Batbaatar et al.

- We can compare the results with those of Batbaatar et al. (51.1% accuracy) on the CrowdFlower dataset
- The best model had only 36.16% of accuracy on the CrowdFlower dataset
- This difference can be traced to the complexity of the model used by Batbaatar et al. that uses a combination of LSTM and CNN



Conclusion

- Random forests have proven to be easier to train than neural networks
- The work could be improved with more computational resources to test more models and more complex neural networks