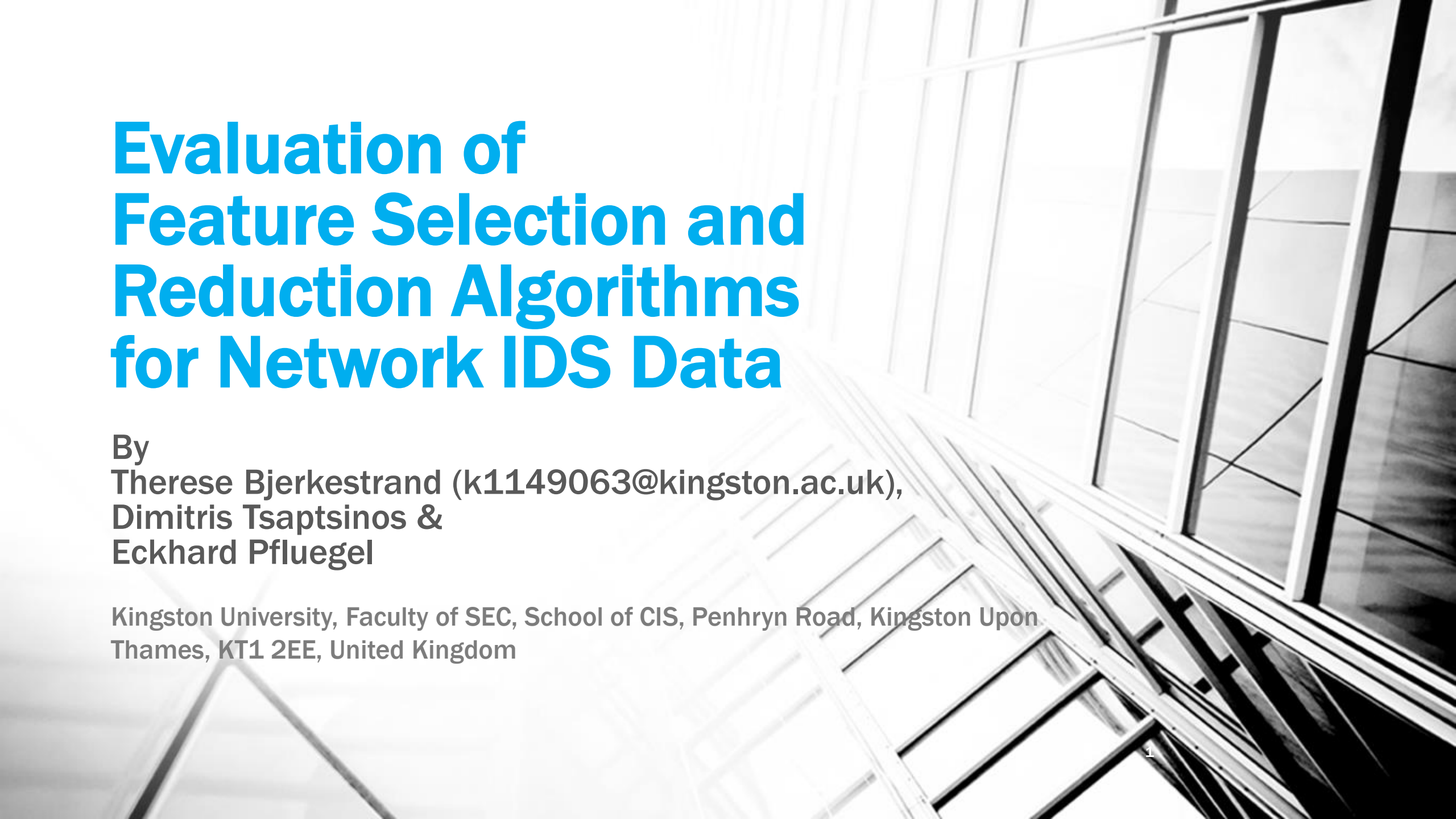
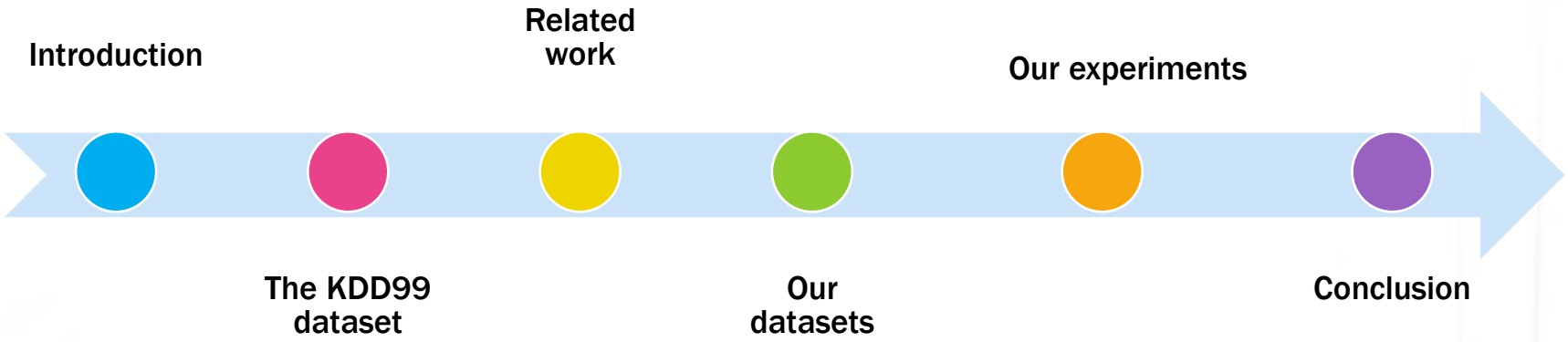
# Evaluation of Feature Selection and Reduction Algorithms for Network IDS Data

By
Therese Bjerkestrand (k1149063@kingston.ac.uk),
Dimitris Tsaptsinos &
Eckhard Pfluegel

Kingston University, Faculty of SEC, School of CIS, Penhryn Road, Kingston Upon Thames, KT1 2EE, United Kingdom



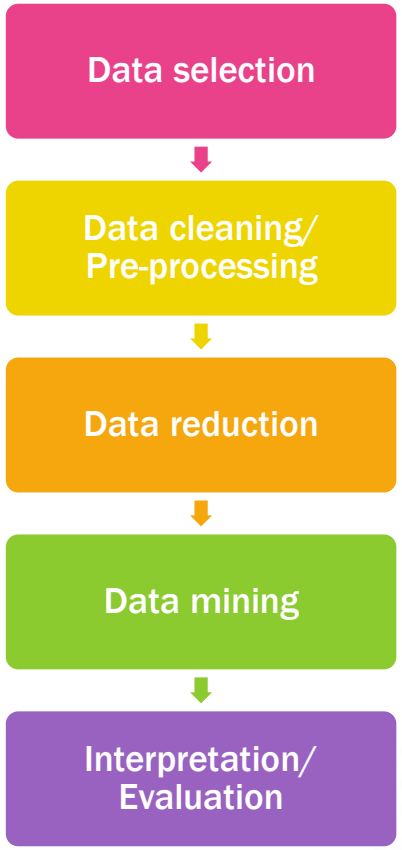
# Content


Introduction
The KDD99 dataset
Related work
Our datasets
Our experiments
Conclusion

# Introduction

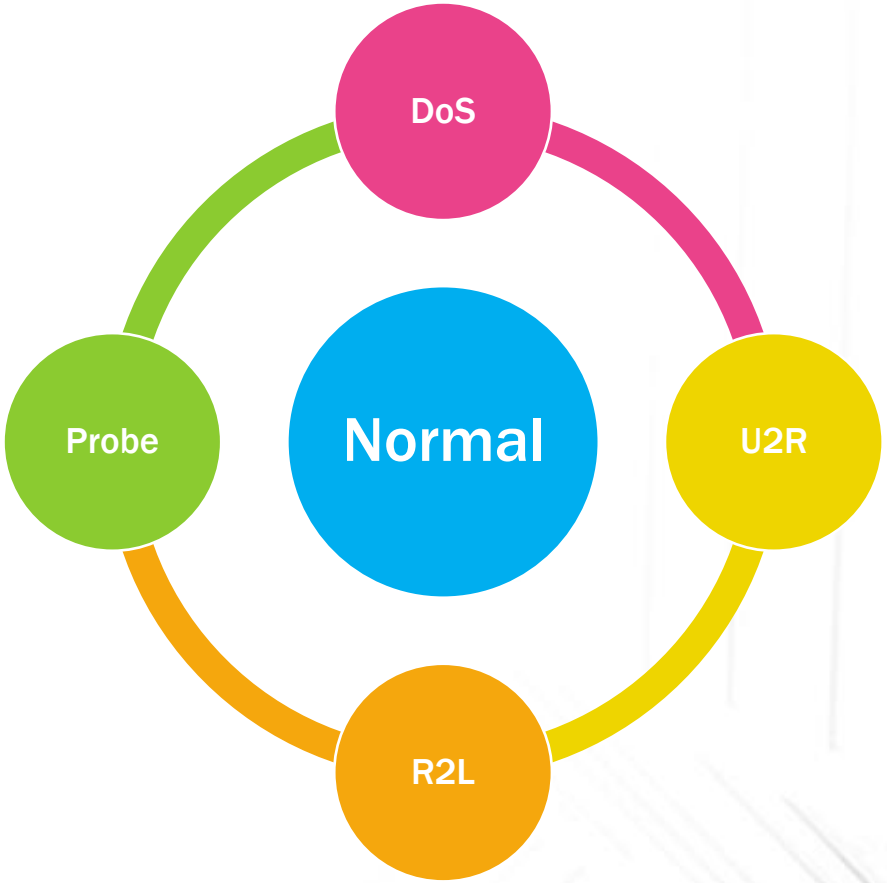
- Knowledge discovery in databases
  - Data reduction
    - Feature selection
  - Data mining

- Intrusion detection
  - Misuse detection
  - Anomaly detection



# The KDD99 Dataset

- Collection of simulated raw TCP dump data collected over a period of nine weeks
- Been particularly exposed to research
- Approx. 5 million instances
- Dataset has 41 attributes
- Instances divided into 4 attack categories and 1 normal category

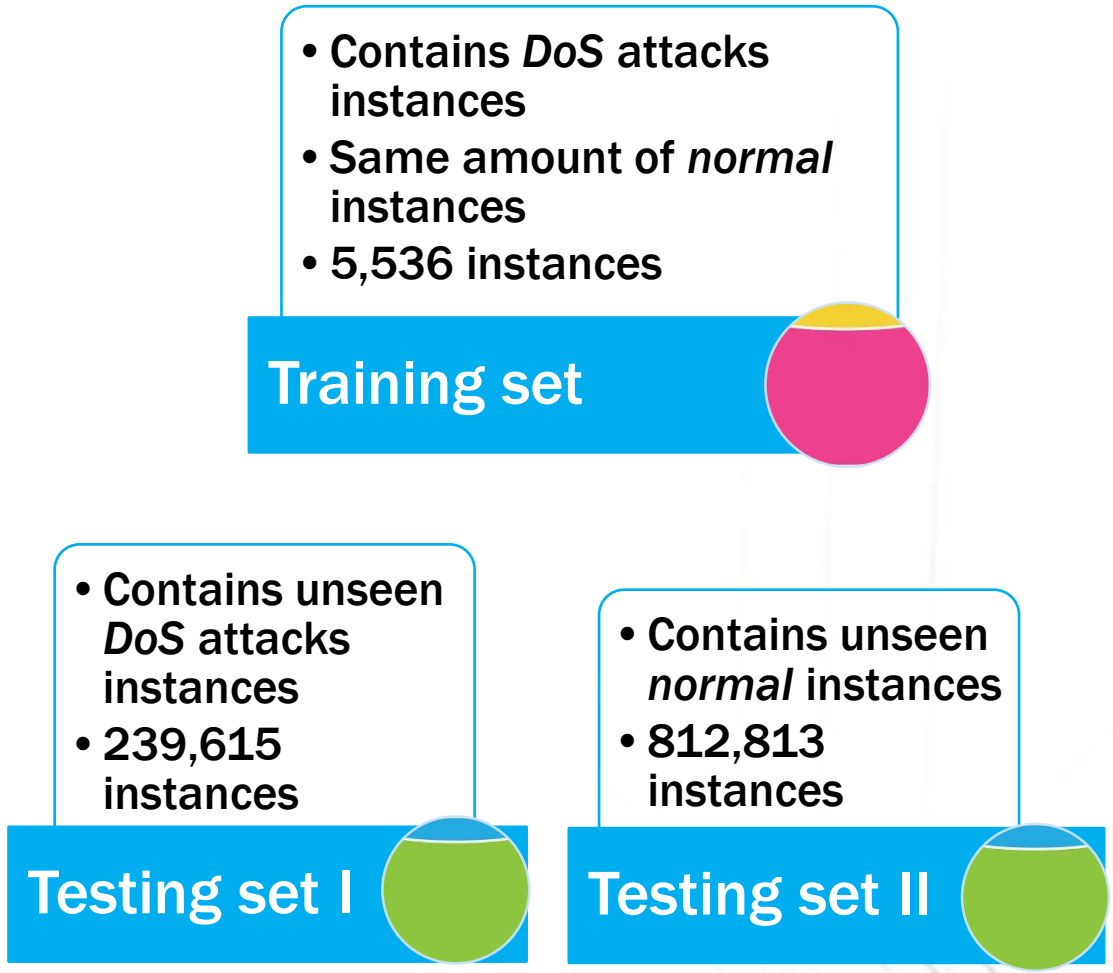


# Related Work

- Panda and Patra
  - Evaluated the performances of ID3, J48 and Naïve Bayes based on 10-fold cross validation test
    - No single best algorithm
    - Naïve Bayes – one of the most effective learning algorithms
    - Decision trees – better for discovering new attacks
- Nguyen and Choi
  - Evaluated performances of a comprehensive set of classifier algorithms, such as Naïve Bayes, J48, JRip and SMO
    - No single algorithm detected all attack categories with a high probability of detection and low false alarm rate
- Ektefa *et al*
  - Used KDD99 dataset and J48 and SVM algorithms for intrusion detection
    - J48 – better performance than SVM in both detection and false alarm rate

# Our Datasets

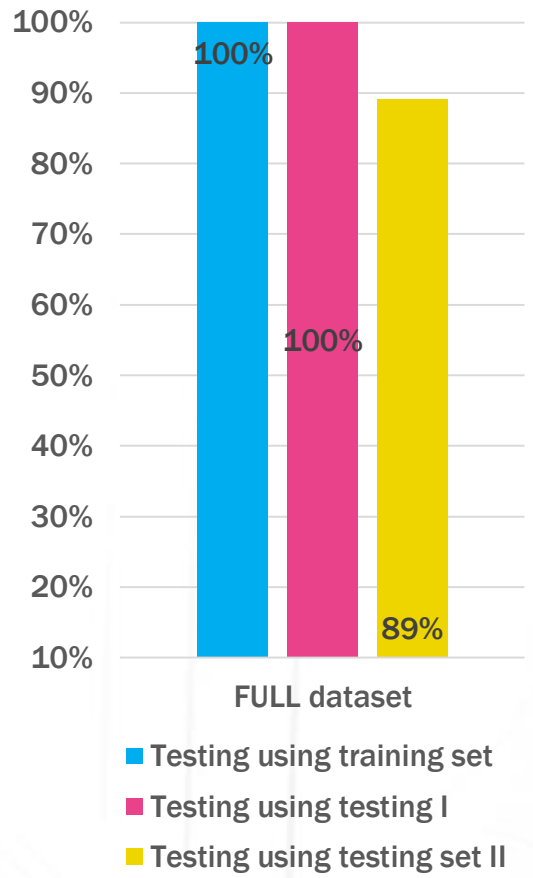
- Three new dataset created
- Only *DoS* and *normal* instances
- Consists of randomly selected instances from the original KDD99 dataset
- No changes made to attributes or class
- Duplicates removed
- Attacks labelled as *DoS* to create common label



# Our Experiments

- Decision tree learning algorithm C4.5 (known as J48 in Weka) was applied to the training set to test the accuracy of J48 before feature selection and reduction algorithms were applied
- Results showed 100% correct classification when using full training set
- Three feature selection algorithms (FSAs) were then chosen and applied to the training set.

| No. | Attribute Evaluator | Search Method |
|---|---|---|
| FSA 1 | Fuzzy Rough Subset Evaluation | Hill Climber |
| FSA 2 | Correlation Attribute Evaluation | Ranker |
| FSA 3 | Cfs Subset Evaluation | Best First |



# Feature Selection Algorithm Results

## Feature Selection Algorithms 1 & 3

- FSA 1 and FSA 3 both produced subsets
- FSA 1 selected only three features: *service, flag* and *src_bytes*
- FSA 3 selected four features: *src_bytes, dst_bytes, dst_host_srv_diff_host* and *dst_host_rerror_rate.*

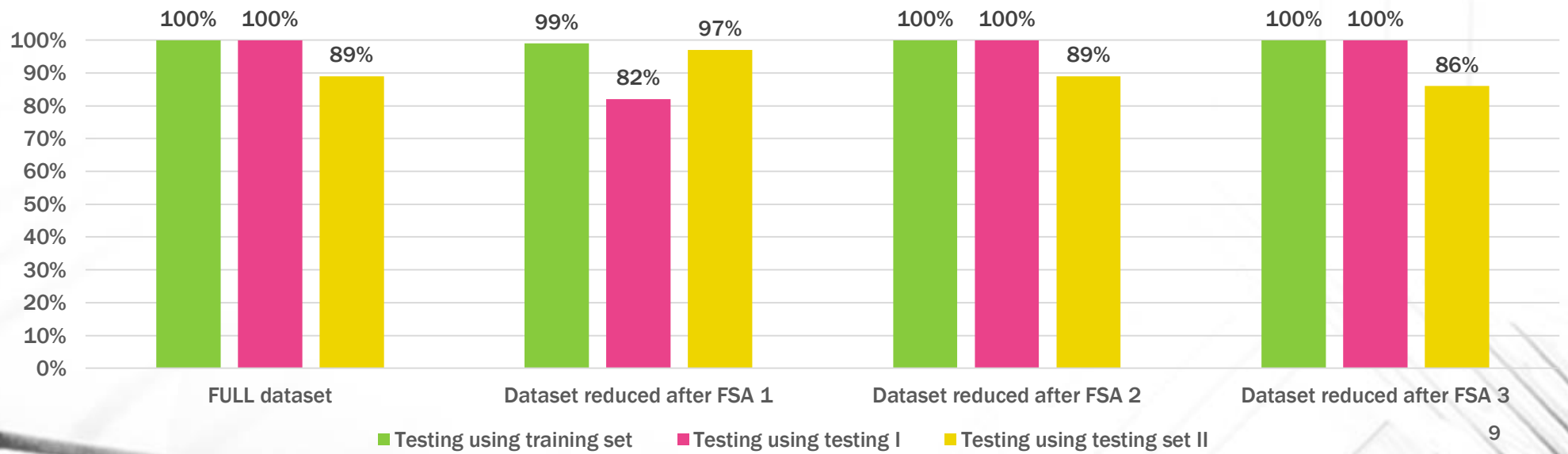
## Feature Selection Algorithm 2

- FSA 2 produced a list, ranking all features from 0-1
- Features ranked as zero were excluded
- 8 features were ranked as zero

# Accuracy Results

- Datasets were reduced according to results of the FSAs.
- J48 performance when using the training set to train the algorithm was stable
- Introducing testing set I and II produced more unstable results

| Dataset | J48 Performance using training set | J48 Performance - testing set I (unseen dos) | J48 Performance - testing set II (unseen normal) |
|---|---|---|---|
| FULL | 100% | 100% | 89% |
| FSA1 | 99% | 82% | 97% |
| FSA2 | 100% | 100% | 89% |
| FSA3 | 100% | 100% | 86% |



# Future Work

- More tests needs to be performed
- Pre-processing of datasets?
- Different learning algorithms
- Different feature selection and reduction algorithms
  - Various attribute evaluators and search methods

# Conclusion

- Feature Selection did not have a great effect on accuracy
- Model is instance dependent
  - Characteristics of data has an effect
- Time taken to build model reduced
  - Reduction based on FSA1 lead to time taken to build model being reduced by 10 seconds while learning algorithm accuracy increased by 8% when testing with unseen *normal* instances