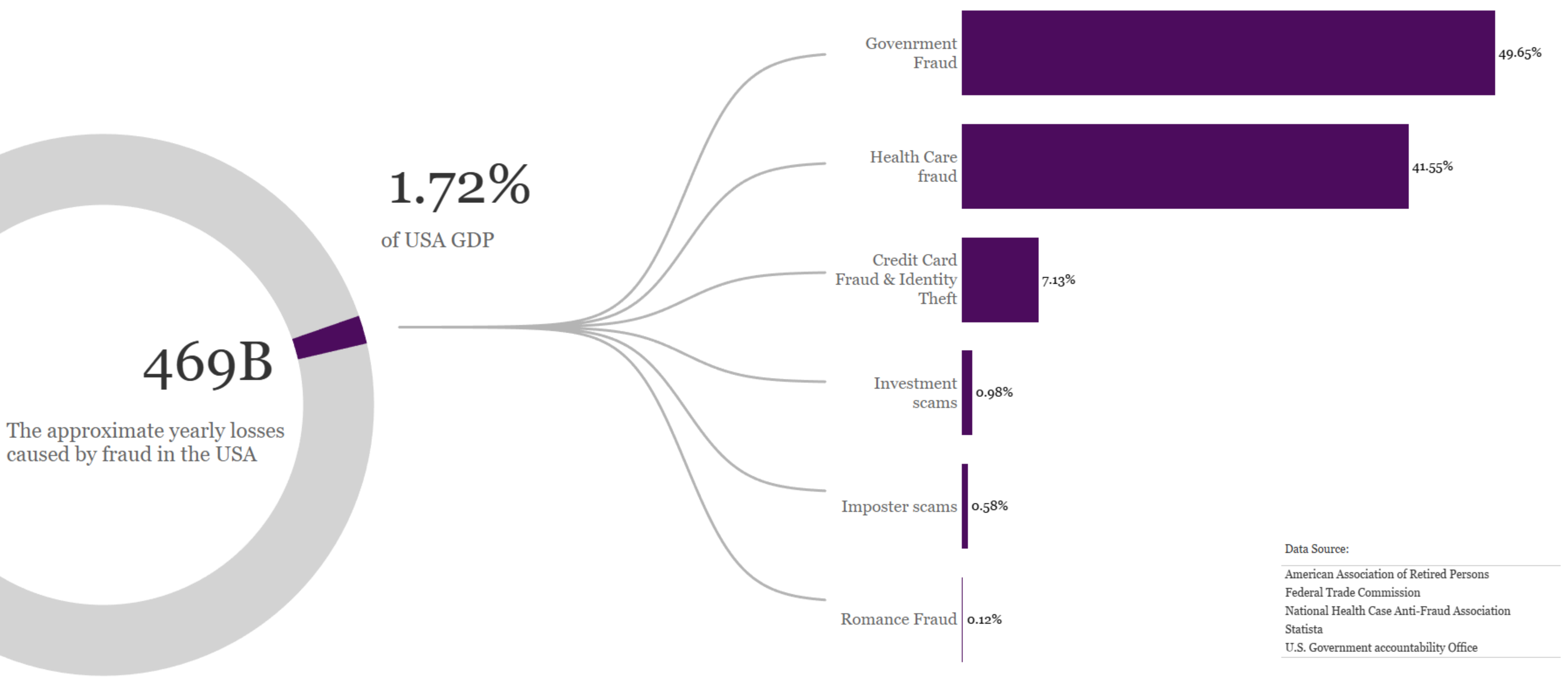
Highly imbalanced data case: Pattern - Guided Feature Selection to Detect Financial Fraud

ID II-5

Dalia Breskuviene and Gintautas Dzemyda,
Data Science and Digital Technologies Institute, Vilnius University



Data Source:
American Association of Retired Persons
Federal Trade Commission
National Health Case Anti-Fraud Association
Statista
U.S. Government accountability Office

Challenges with Imbalanced Data

In financial fraud detection, datasets are typically highly imbalanced, with fraudulent transactions making up only a small fraction compared to legitimate ones. This imbalance creates significant challenges, as traditional machine learning models often lean toward the majority class, leading to poor detection of fraudulent cases. Rapid response time is also crucial in this context; detecting fraud in real-time or near real-time can prevent further financial loss and limit the impact on affected accounts.

Pattern-Guided Feature Selection

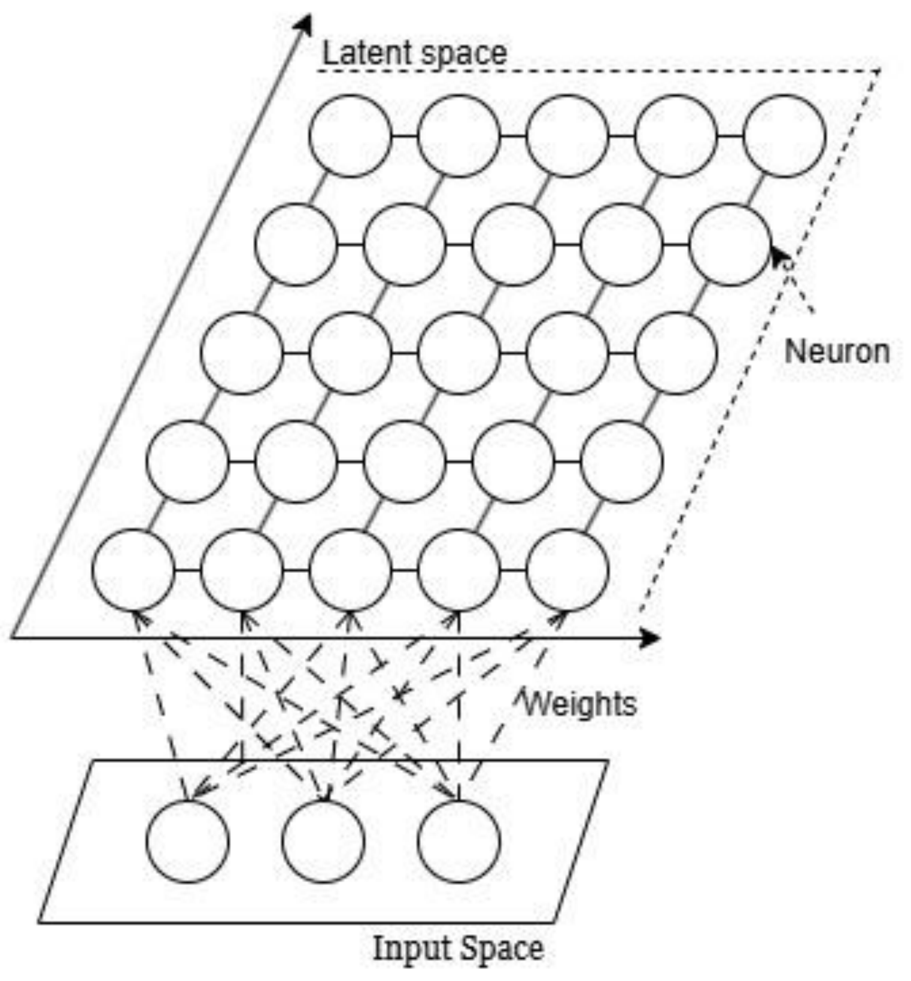
Pattern-guided feature selection leverages identifiable patterns within fraudulent transactions to enhance the detection process. Unlike traditional feature selection techniques that rely solely on statistical measures, this approach incorporates behavioral patterns to prioritize features most relevant for distinguishing fraud.

Proposed Method

Algorithm FID-SOM (Feature selection for imbalanced Data Using SOM)

```
Require X: Dataset
Require params: SOM parameters
Require d: Desired number of features.
Ensure: features subset ensuring high classifier performance
Procedure SELECTFEATURES:
1. train SOM using parameters params with dataset X
2. form a new dataset $W_{BMU}$ containing $n_{BMU}$ weight vectors of m
   attributes corresponding to m features of dataset X
3. normalize $W_{BMU}$ dataset attributes to a scale of [0,1]
4. calculate the variance of each attribute
5. sort attributes based on variance in descending order
6. select d attributes from the top of the list
7. select features for dataset X corresponding to the kept attributes
```

We propose employing Self-Organizing Maps (SOM), which is a special type of artificial neural network, to select valuable features. The SOM algorithm is applied to the dataset to cluster similar transactions together and identify patterns of fraudulent behavior.



This method dynamically adapts to the data's inherent characteristics, ensuring an automatic and data-driven feature selection process. It significantly enhances the method's suitability for diverse scientific applications, where datasets often vary in dimensionality and complexity.

FID-SOM was compared with univariate feature selection methods utilizing the F-test, χ^2-test and mutual information, the Recursive Feature Elimination method, and the XGB Importance method. The goodness of the feature selection methods was evaluated using F1 score, MCC, G-Mean, AUC-PR, and AUC-ROC metrics when performing XGBoost, CatBoost, and Random algorithms on three datasets.

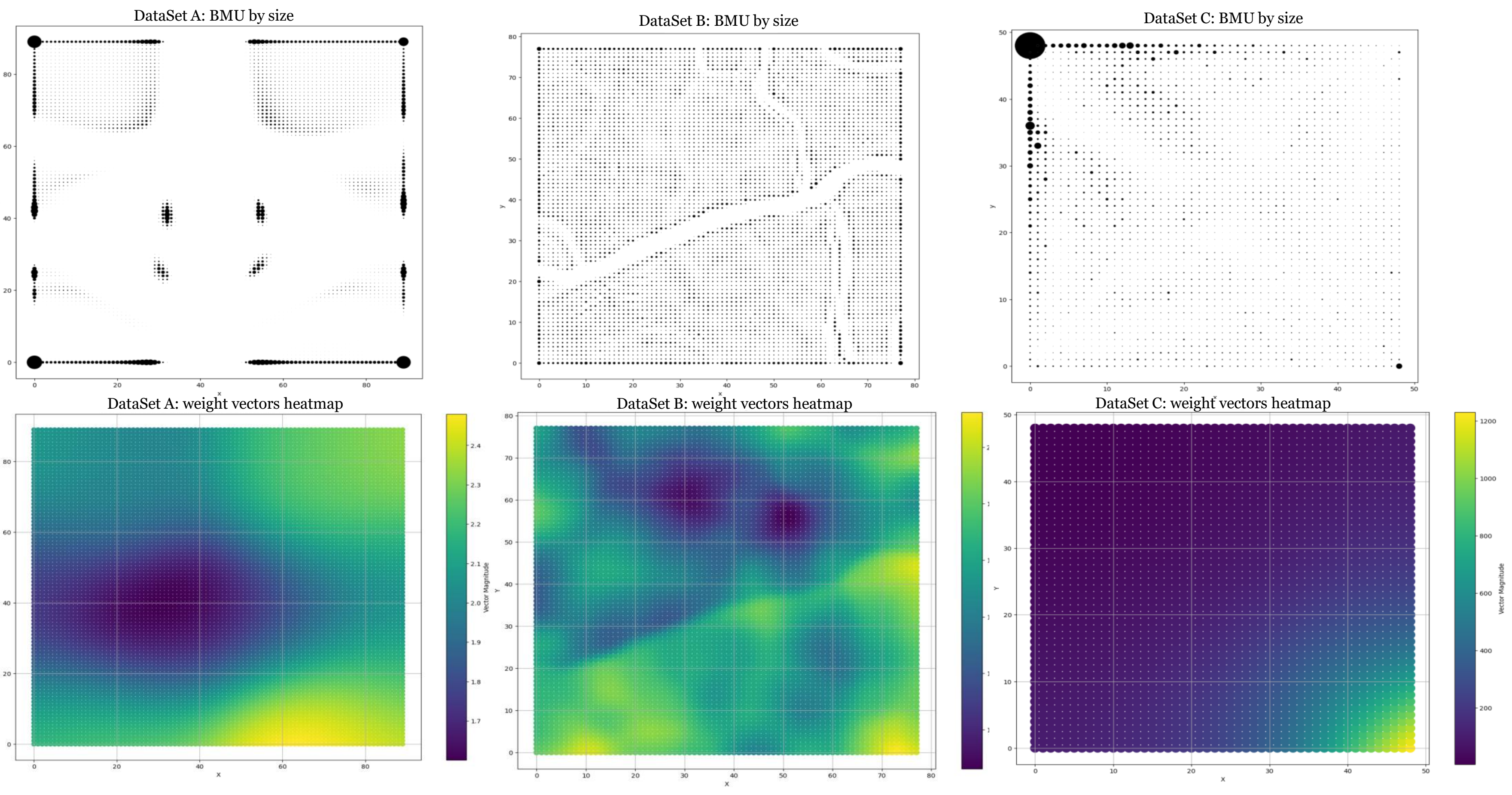
Experimental Results

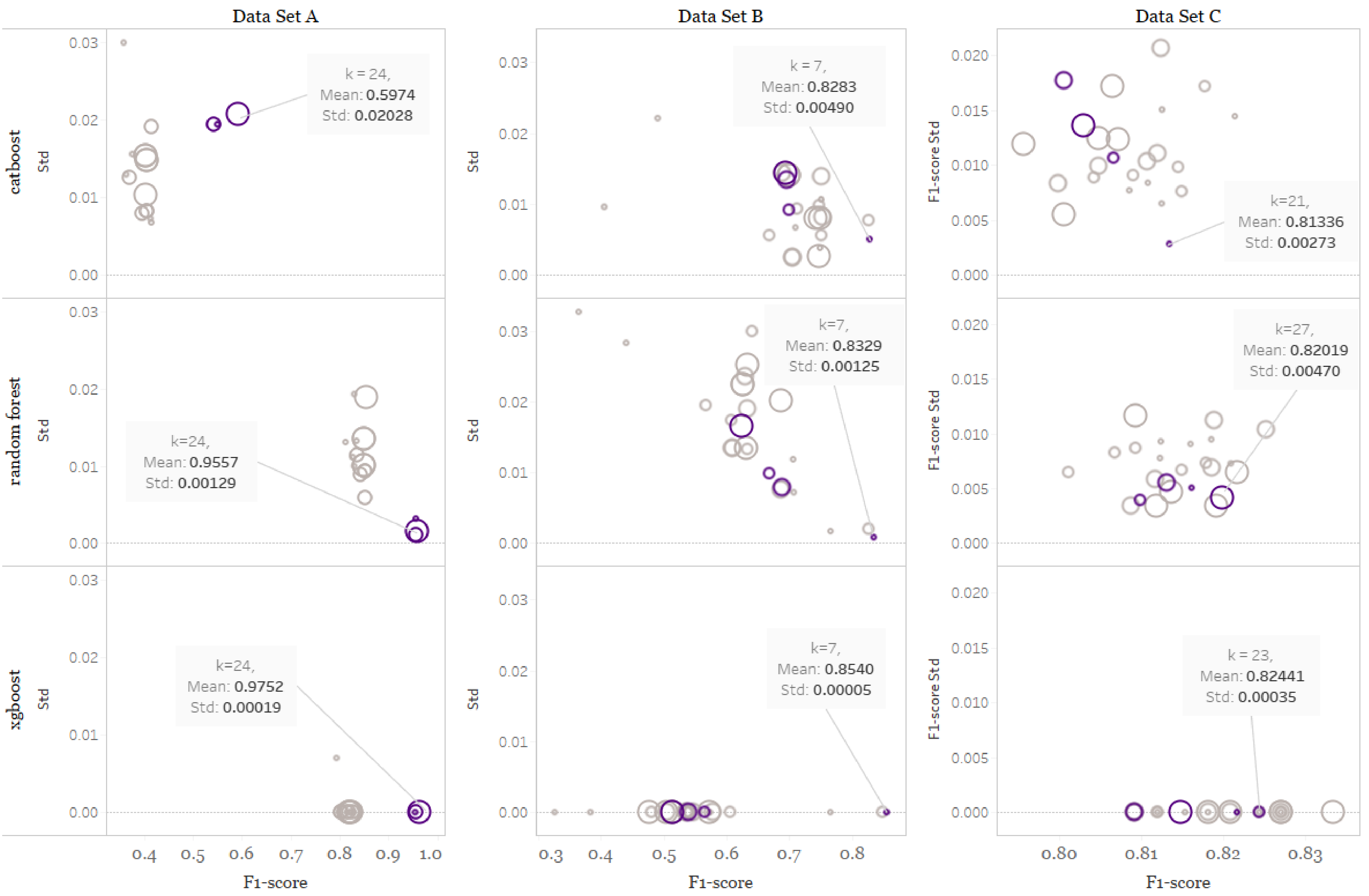
Category	DataSet-A	DataSet-B	DataSet-C
Not Fraud (Percentage)	99.86%	99.48%	99.83%
Fraud (Percentage)	0.14%	0.52%	0.17%
# of instances	3,445,553	1,852,394	284,807
# of features	25	11	29

Each dataset is split using a time-based approach. The earliest 80% of instances are used for training, and the rest of the data, which has timestamps later than the training set, is left for testing.

We have used the categorical feature encoding method, the James-Stein encoder, and discovered it as comparatively best for imbalanced data in the paper [2], where six feature encoders were compared.

Trained SOM Visualization





The results suggest that FIDSOM is a consistently high-performing feature selection method across all datasets, demonstrating reliable outcomes compared to alternative approaches. While other methods achieve moderate performance, they generally are less universal and reach the levels observed with FIDSOM only in specific datasets. These findings indicate that FIDSOM may provide a more effective approach to feature selection in this context, though further evaluation may be necessary to confirm its advantages across broader scenarios.

References

1. "D. Breskuvienė, G. Dzemyda (2023). Imbalanced data classification approach based on clustered training set. In: Dzemyda G., Bernatavičienė J., Kacprzyk J. (Eds.), Data Science in Applications. Studies in Computational Intelligence. Springer"
2. "Breskuvien e, D., Dzemyda, G.: Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. International Journal of Computers Communications & Control 18(3) (2023)"