STable Ensemble Feature-Ranking (STEF-Rank)

CSI5388 - Group 4

Group 4 Members

AS

Alexandra Sklokin

AG

Aryan Gulati

PP

Pranav Pawar

AD

Anisha Deochake

Table of Contents

Background

Motivation

Literary Review

STEF-Rank

01

Motivation

- **Feature Selection**:
  - is a critical step in Machine Learning, involving the identification and removal of irrelevant or redundant features.
  - enhances the data preprocessing phase, contributing to the development of more effective models.
  - is pivotal in identifying relevant features for effective threat detection in cybersecurity.

- Cybersecurity datasets, such as network logs, often contain numerous irrelevant or noisy features.

Literary Review

- **Filter Methods**:
  - valuable for reducing dimensionality and enhancing intrusion detection model performance.
  - utilize some criteria and threshold to select features.

Figure 1. Filter Methods [1]

Literary Review

Prominent Filter Methods:

- **Mutual Information:**
  - Measures statistical dependence between two variables.
  - Indicates the amount of information shared between the feature and the target variable.
- **Variance Threshold:**
  - Filters features based on their variance.
  - Useful for identifying features with low variability.
- **SelectKBest:**
  - Selects the top k most important features based on statistical tests like chi-squared, ANOVA F-test, and mutual information.

Literary Review

- **Wrapper Methods**:
  - measure feature importance based on usefulness during ML model training.
  - are computationally more expensive but offer benefits such as interacting with the classifier and a more comprehensive search of feature space.

Figure 2. Wrapper Methods [1]

Literary Review

Prominent Wrapper Methods:

- **Backward Elimination:**
  - Iterative method starting with all features and removing the least significant feature at each iteration.
- **Recursive Feature Elimination (RFE)**:
  - Wrapper-like greedy optimization algorithm that aims to find the best-performing feature subset.

Literary Review

Limitations of Traditional Feature Selection Techniques:

Filter Methods:

- *Sensitivity to Thresholds:* Arbitrary threshold choices impact outcomes.
- *Limited Consideration of Dependencies:* Overlooks inter-feature dependencies.
- *May Exclude Relevant Features:* Strict criteria risk excluding contextually important features.

Wrapper Methods:

- *Computational Expense:* More resource-intensive than filter methods.
- *Possible Overfitting:* Iterative optimization may lead to overfitting.
- *Limited Interpretability:* Increased complexity challenges interpretability.

STable **E**nsemble **F**eature-**Rank**

- **Ensemble Approach:**
  - STEF-Rank employs a **bagging ensemble** technique for Feature Selection.
  - Multiple 'weak' Feature Selection techniques contribute to the ensemble.

- **Stability Through Resampling:**
  - Resampling is a key element, creating subsets of the dataset for stability.
  - Each 'weak' feature selector is applied multiple times on the same dataset, using these resampled subsets.

- **Ranking Process**:
  - Features are given ranks based on their performance in resampling and 'weak' Feature Selection.
  - The rank matrix provides insights into the importance of each feature.

STEF-Rank

- **Ensemble Ranking:**
  - The ensemble aggregates individual rankings, producing an overall **STEF-Rank** for each feature.
  - This process enhances the robustness of feature selection.

- **Thresholding for Selection:**
  - A threshold is applied to filter out features, selecting only those with a ranking above a specified value (e.g., 0.5).
  - This step helps focus on the most important features.

STEF-Rank

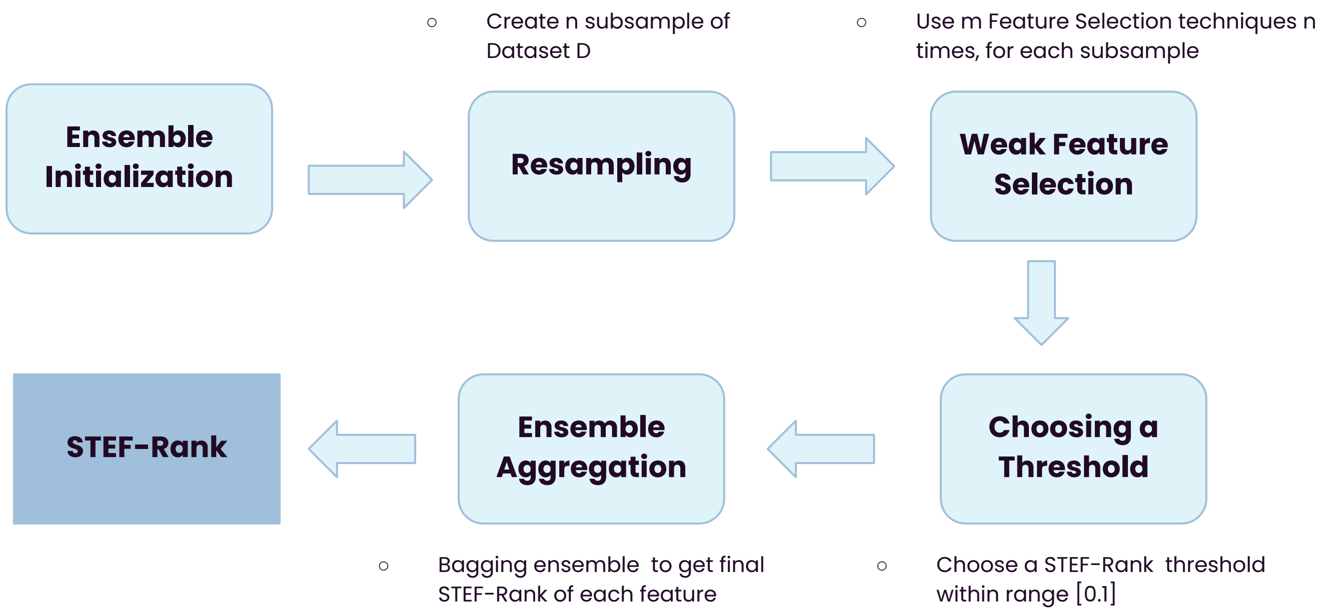


Figure 3. STEF-Rank Pipeline

STEF-Rank

Algorithm 1 STEF-Rank (D, FS, n, $threshold$)

Input: dataset D; set of Feature Selection techniques FS; number of resamplings n; feature selection $threshold$

for each FS_i in FS **do**

  $SubD \leftarrow Resampling(D, n)$ # create n subsets using resampling

  for each $SubD_j$ in $SubD$ **do**

    $feature\_selection_j \leftarrow FS_i(SubD_j)$ # apply 'weak' feature selector to resampling

  end for

  for each $feature$ in $D.features$ **do**

    $feature\_rank_i[feature] \leftarrow CountOccurances(feature\_selection, feature) \div (n * m)$ # rank matrix for the 'weak' learner

  end for

end for

for each $feature$ in $D.features$ **do**

  $ensemble\_ranking[feature] \leftarrow Sum(feature\_rank[feature])$ # ensemble ranking

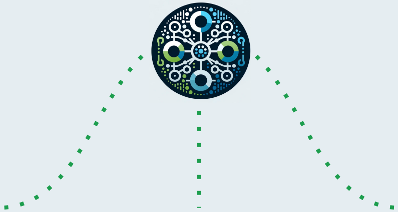
end for

$best\_features \leftarrow ensemble\_ranking.iloc(ranking \geq threshold)$ # features with high rank

return best_features # Output: dictionary with features as keys and rankings as values

02

Experimental Setup



Datasets Methodology Baseline & Metrics

Dataset

We will be investigate **five** datasets, for **DDoS attack classification**, generated from the following two simulated attack scenarios:

1. The **CSE CIC IDS2018** dataset, provided by the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC) focuses on intrusion detection in 2018.

   DDOS Attack Type: *PORTMAP*

2. The **CIC DDoS2019** dataset, from the Canadian Institute for Cybersecurity (CIC) specifically targets Distributed Denial of Service (DDoS) attacks in 2019.

   DDOS Attack Types: *UDP, LDAP, SYN, and NETBIOS*

Dataset

Datasets	UDP	LDAP	NETBIOS	PORTMAP	SYN
Attack/Benign Samples	3134 / 3083	200 000 / 5053	200 000 / 1687	128 027 / 97 718	100 000 / 381
Number of Features	78	87	87	87	87

Table 1. Dataset Distributions

Methodology

1. **Data Preprocessing :** Data cleaning, removing time-dependent variables such as Timestamp, and normalization.
2. **Train/Test Split:** Split the dataset 70/30.
3. **Baseline and Novel Feature Selection :** Using traditional filter/wrapper methods as baselines, and the STEF- Rank technique to general subsets of the feature set.
4. **Classification & Validation :** Perform classification with the various classification models (such as Random Forest & XGBoost) using 10-Fold Cross-Validation on training set.
5. **Evaluation Metrics :** Collect F1-Score, Precision-Recall AUC, for testing set.
6. **Comparison to Baseline :** Tests of statistical significance such as T-test and Wilcoxon Signed Rank, so compare the performance of STEF-Rank to our baselines.

Evaluation Metrics

The following performance metrics are suitable for the **unbalanced classification** problem.

1. The **F1 Score** is a measure of how the model performs, taking into account both precision and recall.

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{1}$$

2. **Precision Recall AUC (Area Under the Curve)** is a metric that assesses the trade off, between precision and recall. It provides a performance value for the model.

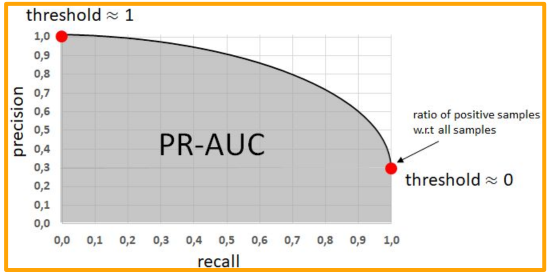


Figure 4. Example Precision Recall Curve

Comparison to Baselines

1. **Friedman Test**: a parametric test used for analyzing randomized complete block designs. Is used to prove/disprove statistical different in **mean.**

2. **Wilcoxon Signed Rank Test**: a parametric test and is used to determine if there's a significant difference between paired groups. Is used to prove/disprove statistical difference in **median.**

Results

Experimental Results Limitations Future Considerations

03

Experimental Results

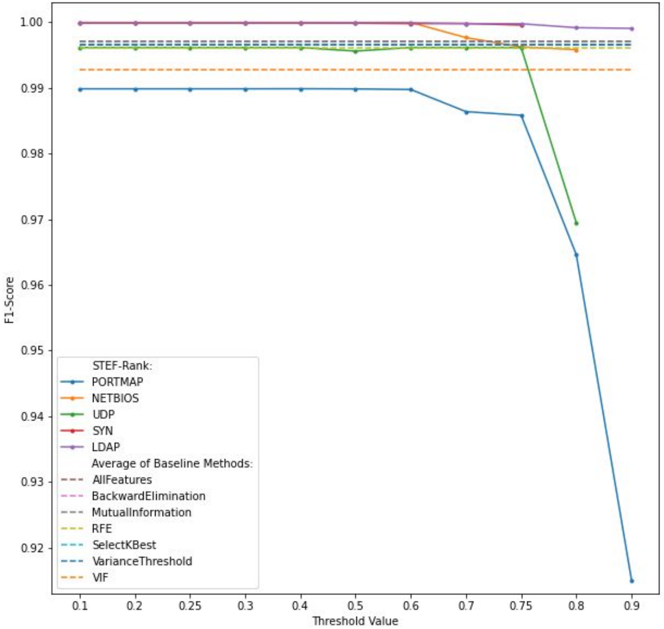


Figure 4. Graph of STEF-Rank and Baseline Methods

- **Performance:** STEF-Rank does not seemingly outperform baseline methods. All models performed exceptionally well, such that results are hard to interpret.
- **Threshold:** when the STEF-Rank threshold is too high, the algorithm removes some relevant features and results in worse performance.
- **AUC-PR:** showed the same relationship between STEF-Rank, baselines, and threshold values.

Experimental Results

Feature Selection Technique	Number of Selected Features	Cross-Validation F1-Score	Testing F1-Score	Testing PR-AUC
AllFeatures	78	**0.9890**	**0.9899**	**0.9997**
BackwardElimination	50	0.9863	0.9874	0.9996
MutualInformation	65	0.9885	0.9896	**0.9997**
RFE	50	0.9858	0.9870	0.9995
SelectKBest	40	0.9855	0.9872	0.9996
VarianceThreshold	36	0.9882	0.9893	**0.9997**
VIF	13	0.9659	0.9705	0.9712
STEF-Rank_0.1	73	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.2	69	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.25	68	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.3	68	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.4	66	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.5	57	**0.9890**	0.9898	**0.9997**
STEF-Rank_0.6	46	0.9889	0.9898	**0.9997**
STEF-Rank_0.7	34	0.9849	0.9864	0.9989
STEF-Rank_0.75	25	0.9843	0.9858	0.9988
STEF-Rank_0.8	14	0.9583	0.9646	0.9878
STEF-Rank_0.9	2	0.8932	0.9150	0.5946

Table 2. Experimental Results for PORTMAP Dataset

For each of the 5 datasets (*PORTMAP, UDP, SYN, LDAP, NETBIOS*):

T-Test:

- The difference in the **mean** F1 and **mean** AUC-PR cannot be proved to be statistically significant. STEF-Rank may not outperform any of the baseline methods.

Wilcoxon:

- The difference in the **median** F1 and **median** AUC-PR cannot be proved to be statistically significant. STEF-Rank may not outperform any of the baseline methods.

Experimental Results

Feature Selection Technique	Number of Selected Features	Cross-Validation F1-Score	Testing F1-Score	Testing PR-AUC
AllFeatures	78	**0.9890**	**0.9899**	**0.9997**
BackwardElimination	50	0.9863	0.9874	0.9996
MutualInformation	65	0.9885	0.9896	**0.9997**
RFE	50	0.9858	0.9870	0.9995
SelectKBest	40	0.9855	0.9872	0.9996
VarianceThreshold	36	0.9882	0.9893	**0.9997**
VIF	13	0.9659	0.9705	0.9712
STEF-Rank_0.1	73	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.2	69	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.25	68	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.3	68	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.4	66	**0.9890**	**0.9899**	**0.9997**
STEF-Rank_0.5	57	**0.9890**	0.9898	**0.9997**
STEF-Rank_0.6	46	0.9889	0.9898	**0.9997**
STEF-Rank_0.7	34	0.9849	0.9864	0.9989
STEF-Rank_0.75	25	0.9843	0.9858	0.9988
STEF-Rank_0.8	14	0.9583	0.9646	0.9878
STEF-Rank_0.9	2	0.8932	0.9150	0.5946

PORTMAP

Feature Selection Technique	Number of Selected Features	Cross-Validation F1-Score	Testing F1-Score	Testing PR-AUC
AllFeatures	85	0.9999	1.0000	1.0000
BackwardElimination	50	0.9998	1.0000	1.0000
MutualInformation	57	0.9999	1.0000	1.0000
RFE	50	0.9996	0.9999	1.0000
SelectKBest	40	0.9997	0.9999	1.0000
VarianceThreshold	7	0.9983	0.9993	1.0000
VIF	15	0.9995	0.9998	1.0000
STEF-Rank_0.1	74	0.9999	1.0000	1.0000
STEF-Rank_0.2	72	0.9999	1.0000	1.0000
STEF-Rank_0.25	72	0.9999	1.0000	1.0000
STEF-Rank_0.3	71	0.9999	1.0000	1.0000
STEF-Rank_0.4	67	0.9999	1.0000	1.0000
STEF-Rank_0.5	54	0.9999	1.0000	1.0000
STEF-Rank_0.6	38	0.9998	1.0000	1.0000
STEF-Rank_0.7	15	0.9946	0.9977	0.9989
STEF-Rank_0.75	9	0.9900	0.9962	1.0000
STEF-Rank_0.8	3	0.9882	0.9958	0.9995
STEF-Rank_0.9		nan	nan	nan

NETBIOS

Feature Selection Technique	Number of Selected Features	Cross-Validation F1-Score	Testing F1-Score	Testing PR-AUC
AllFeatures	85	0.9998	0.9999	1.0000
BackwardElimination	50	0.9997	0.9998	1.0000
MutualInformation	45	0.9997	0.9999	1.0000
RFE	50	0.9998	0.9998	1.0000
SelectKBest	40	0.9997	0.9997	1.0000
VarianceThreshold	10	0.9948	0.9984	0.9996
VIF	17	0.9996	0.9998	1.0000
STEF-Rank_0.1	82	0.9998	0.9999	1.0000
STEF-Rank_0.2	74	0.9998	0.9999	1.0000
STEF-Rank_0.25	71	0.9998	0.9999	1.0000
STEF-Rank_0.3	64	0.9998	0.9999	1.0000
STEF-Rank_0.4	52	0.9998	0.9999	1.0000
STEF-Rank_0.5	39	0.9998	0.9999	1.0000
STEF-Rank_0.6	27	0.9997	0.9998	1.0000
STEF-Rank_0.7	13	0.9996	0.9998	1.0000
STEF-Rank_0.75	6	0.9992	0.9996	0.9999
STEF-Rank_0.8		nan	nan	nan
STEF-Rank_0.9		nan	nan	nan

SYN

Feature Selection Technique	Number of Selected Features	Cross-Validation F1-Score	Testing F1-Score	Testing PR-AUC
AllFeatures	85	0.9999	0.9999	1.0000
BackwardElimination	50	0.9999	0.9999	1.0000
MutualInformation	47	0.9999	0.9999	1.0000
RFE	50	0.9998	0.9998	1.0000
SelectKBest	40	0.9998	0.9999	1.0000
VarianceThreshold	17	0.9999	0.9999	1.0000
VIF	18	0.9992	0.9997	1.0000
STEF-Rank_0.1	75	0.9999	0.9999	1.0000
STEF-Rank_0.2	73	0.9999	0.9999	1.0000
STEF-Rank_0.25	73	0.9999	0.9999	1.0000
STEF-Rank_0.3	73	0.9999	0.9999	1.0000
STEF-Rank_0.4	70	0.9999	0.9999	1.0000
STEF-Rank_0.5	53	0.9999	0.9999	1.0000
STEF-Rank_0.6	44	0.9998	0.9999	1.0000
STEF-Rank_0.7	30	0.9997	0.9998	1.0000
STEF-Rank_0.75	22	0.9996	0.9998	1.0000
STEF-Rank_0.8	4	0.9983	0.9992	0.5001
STEF-Rank_0.9	1	0.9981	0.9991	0.4999

LDAP

Feature Selection Technique	Number of Selected Features	Cross-Validation F1-Score	Testing F1-Score	Testing PR-AUC
AllFeatures	85	0.9984	0.9961	0.9999
BackwardElimination	50	0.9984	0.9961	0.9999
MutualInformation	59	0.9984	0.9961	0.9999
RFE	50	0.9984	0.9940	0.9999
SelectKBest	40	0.9986	0.9961	0.9999
VarianceThreshold	37	0.9984	0.9961	0.9999
VIF	20	0.9979	0.9940	1.0000
STEF-Rank_0.1	73	0.9984	0.9961	0.9999
STEF-Rank_0.2	73	0.9984	0.9961	0.9999
STEF-Rank_0.25	72	0.9984	0.9961	0.9999
STEF-Rank_0.3	71	0.9984	0.9961	0.9999
STEF-Rank_0.4	70	0.9984	0.9961	0.9999
STEF-Rank_0.5	63	0.9984	0.9956	0.9999
STEF-Rank_0.6	45	0.9984	0.9961	1.0000
STEF-Rank_0.7	29	0.9975	0.9961	0.9980
STEF-Rank_0.75	23	0.9972	0.9961	0.9989
STEF-Rank_0.8	4	0.9745	0.9694	0.9572
STEF-Rank_0.9		nan	nan	nan

UDP

Table 3. All Experimental Results

* Experiment was also repeated for CSI5388 Assignment 2 and 3 datasets, with similar results. We were unable to obtain a positive results for STEF-Rank.

Limitations

1. **No Proof of Improved Performance**: we were unable to prove that our novel technique showed significant improvement other baselines.
2. **Explainability:** with n feature selection techniques, and m subsamples, this method provides a matrix of results, before obtaining the final aggregated results. Stability techniques also introduce disagreement on Feature Selection in nxm subsets.
3. **Time:** depending on which ‘weak’ Feature Selection techniques were used, this method can take a long time, and may require parallelism for improved performance.
4. **Choice of Threshold:** is unclear and should be done experimentally.

Future Considerations

1. **Choice of 'Weak' Feature Selection Thresholds:** each feature selection technique has a threshold (ex. variance for Variance Threshold, number of features for Backward Elimination, etc.). By selecting stricter threshold, there will be more disagreement between methods.
2. **Choice of 'STEF-Rank' Parameters:** such as the number and selection of weak Feature Selection techniques, the threshold for the final rank, and the number of subsamples of the dataset.
3. **Boosting instead of Bagging:** bagging to sequentially select smaller and smaller subsamples of the feature set.

Future Considerations

4. **Federated Learning:** each 'remote' agent can provide a list of significant features. The global model can use STEF-Rank to return best global features. May introduce improved performance on the 'remote' agents.
5. **Multi-Modal Learning:** scenarios where data is collected from different modes (text, audio, images, etc.).
6. **Online Learning:** rather than using subsamples for stability, use windows of online data streams.
7. Investigate STEF-Rank on more datasets to prove statistical significance. Investigate the effect of different data distributions, to prove effectiveness of stability techniques.

Thank you

Questions?