APPLYING ONE-CLASS LEARNING ALGORITHMS FOR PREDICTING PHAGE-BACTERIA INTERACTIONS

Juan Fernando López[4], Diogo Leite[1], Grégory Resch[3], Yok-Ai Que[2], Xavier Brochet[1] and Carlos Peña[1]

[1] *School of Business and Engineering Vaud (HEIG-VD), University of Applied Sciences Western Switzerland (HES-SO), Switzerland & Swiss Institute of Bioinformatics (SIB)*

[2] *Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland.*

[3] *Department of Intensive Care Medicine, Bern University Hospital (Inselspital), Bern, Switzerland*

[4] *Autonomous University of the West (Universidad Autónoma de Occidente), Cali, Colombia*

Abstract

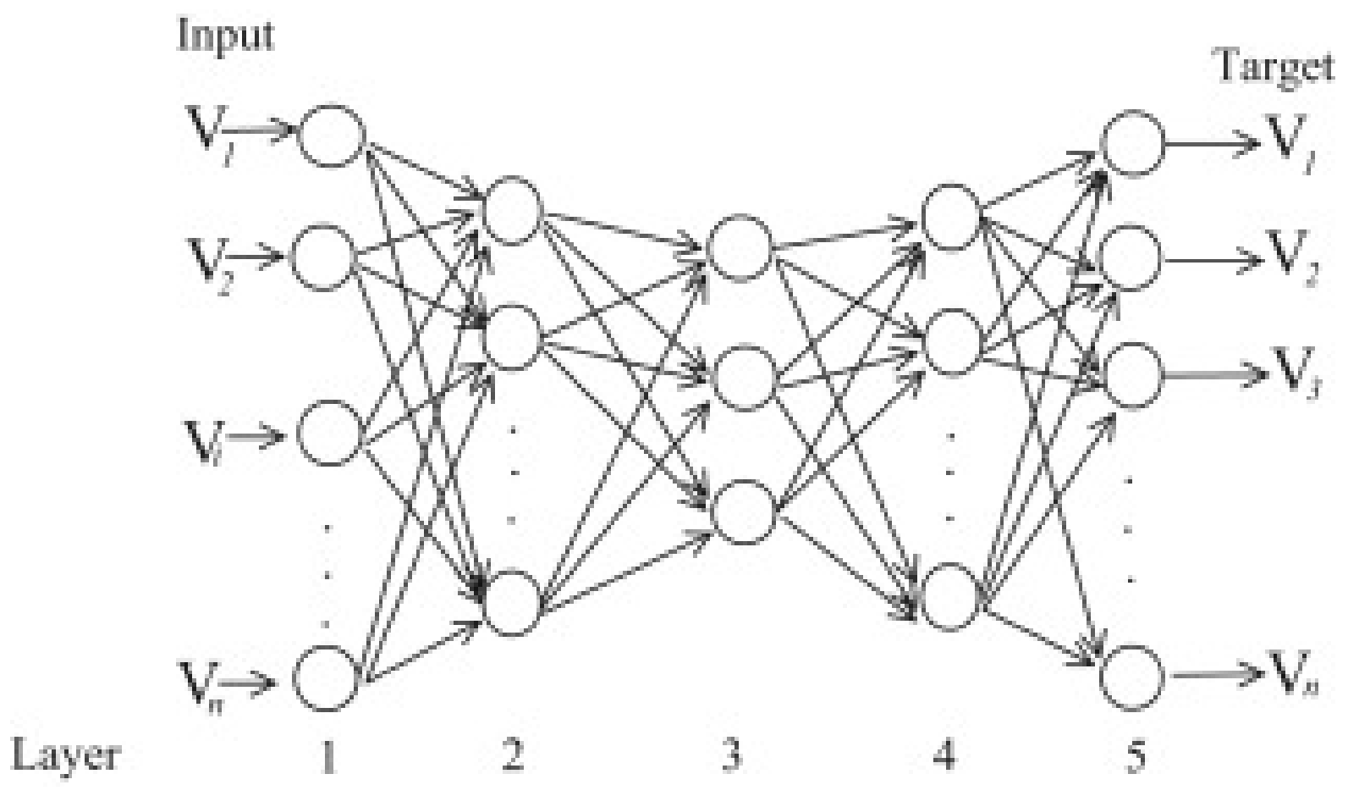
The misuse of antibiotic drugs contributes to the emergence and rapid dissemination of antibiotic resistance worldwide, threatening medical progress. A re-emerging therapy, dubbed phage-therapy, might represent an alternative for this. Phage-therapy is based on bacteriophages that specifically infect and kill bacteria during their life cycle. The success of phage therapy mainly relies on the exact matching between the bacteria and the phage. However, this is a time-consuming process achieved only in laboratories .

Hence, the fast identification of potential phage candidates capable of dealing with a given bacteria is essential for using phage-therapy in routine. Machine learning algorithms constitute a promising approach to achieve this goal. Unfortunately, public databases contain highly imbalanced interaction data (i.e., only positive phage-bacterium interactions); making it harder to use classic machine learning algorithms that needs relatively-balanced classes to work. To address this problem, we are exploring the use of One-Class learning methods, which are robust tools to deal with imbalanced datasets.

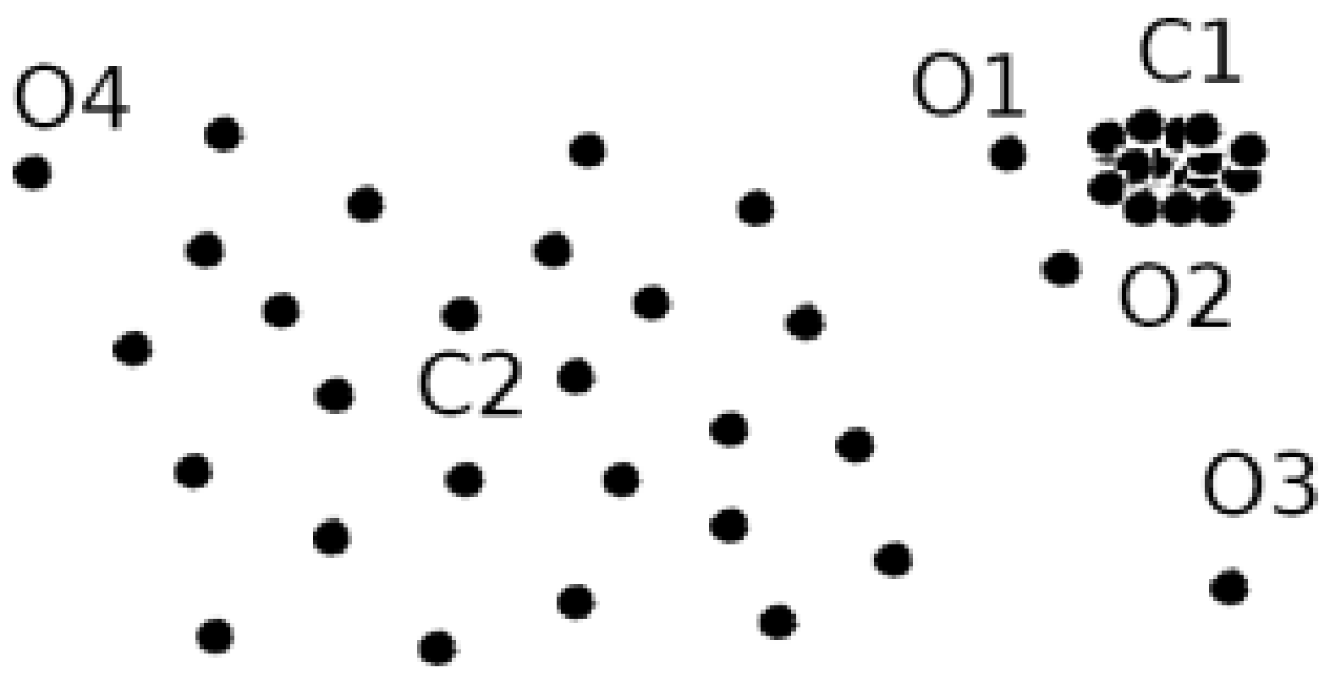
We have tested an odd number of One-Class learning techniques merged with the ensemble-learning paradigm on real phage-bacteria interactions and obtained accurate results in different types of metrics (e.g. accuracy and f1-score up to 80%). Further work could include developing new methods for One-Class classification and applying them to other types of real data.

One-Class Learning Algorithms (Ensemble)

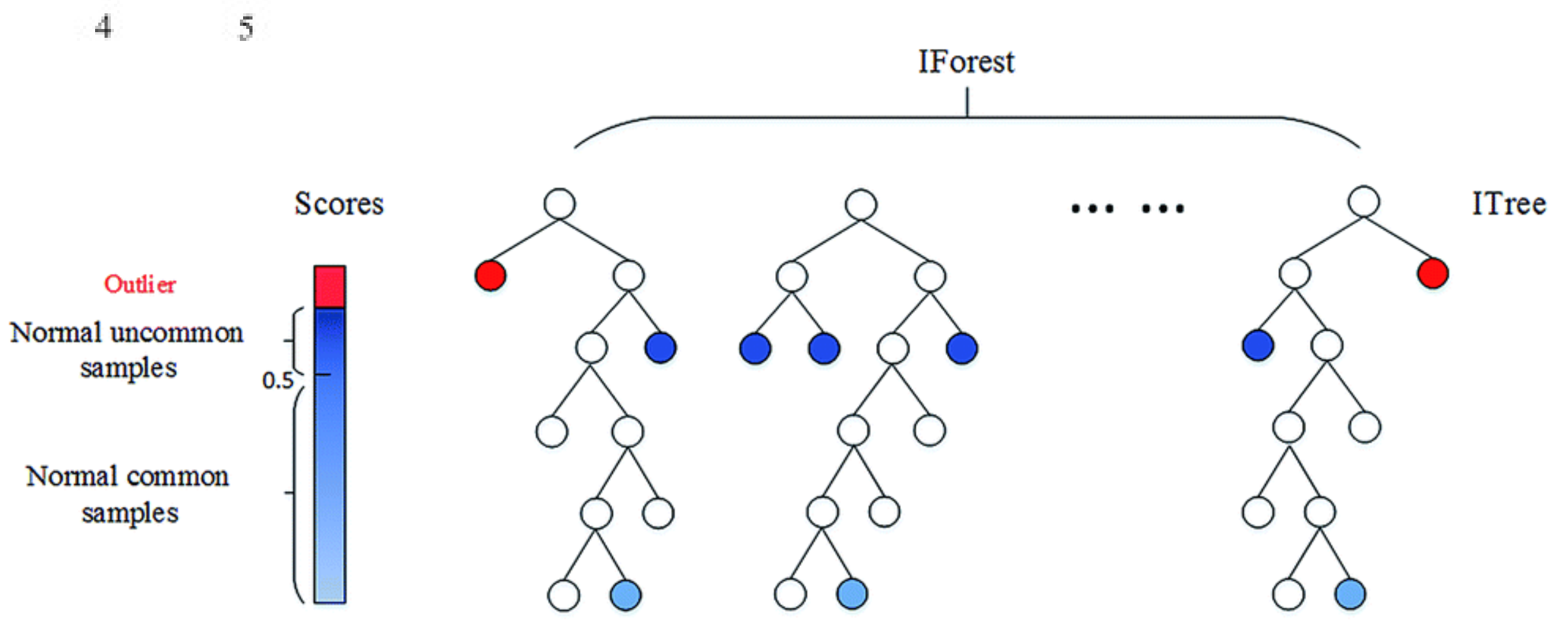
Replicator Neural Network (RNN)



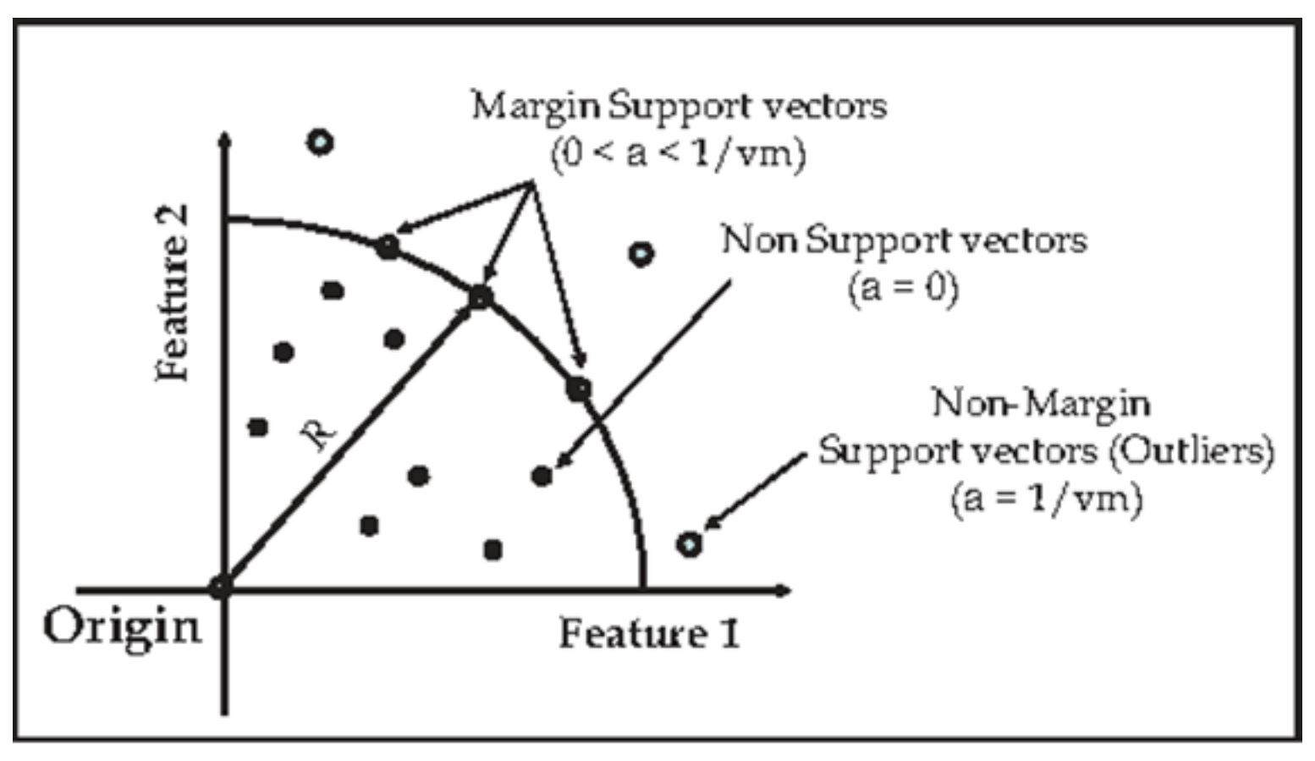
Local Outlier Factor (LOF)



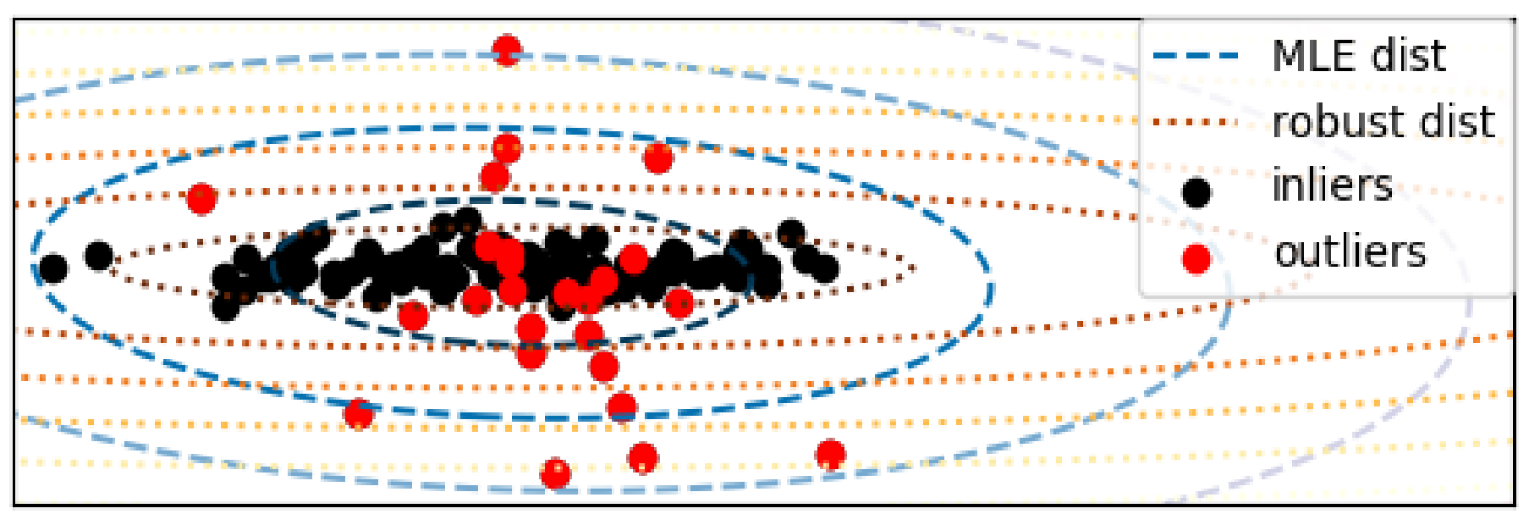
Isolation Forest (iForest)

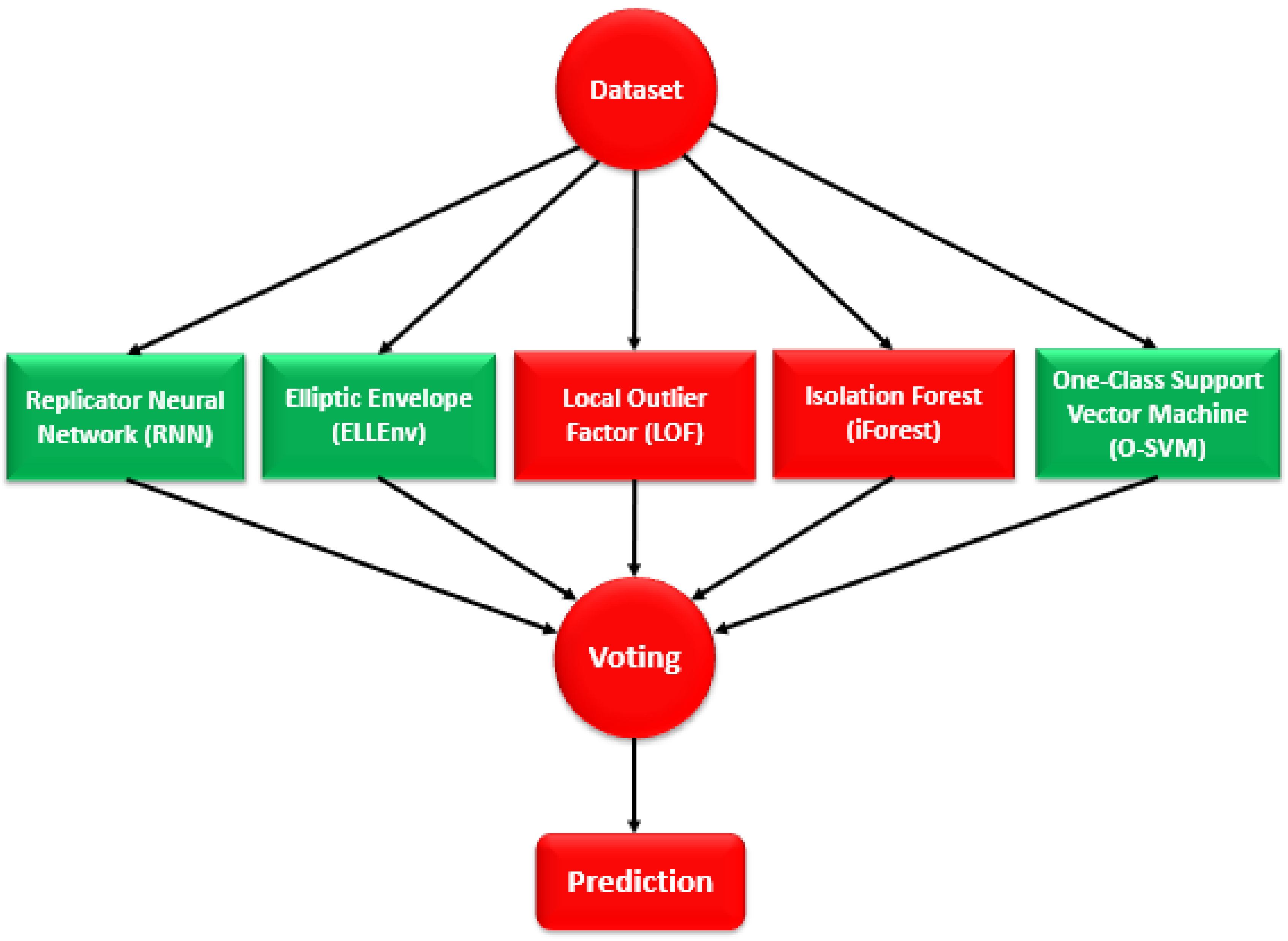


One-Class Support Vector Machine (O-SVM)

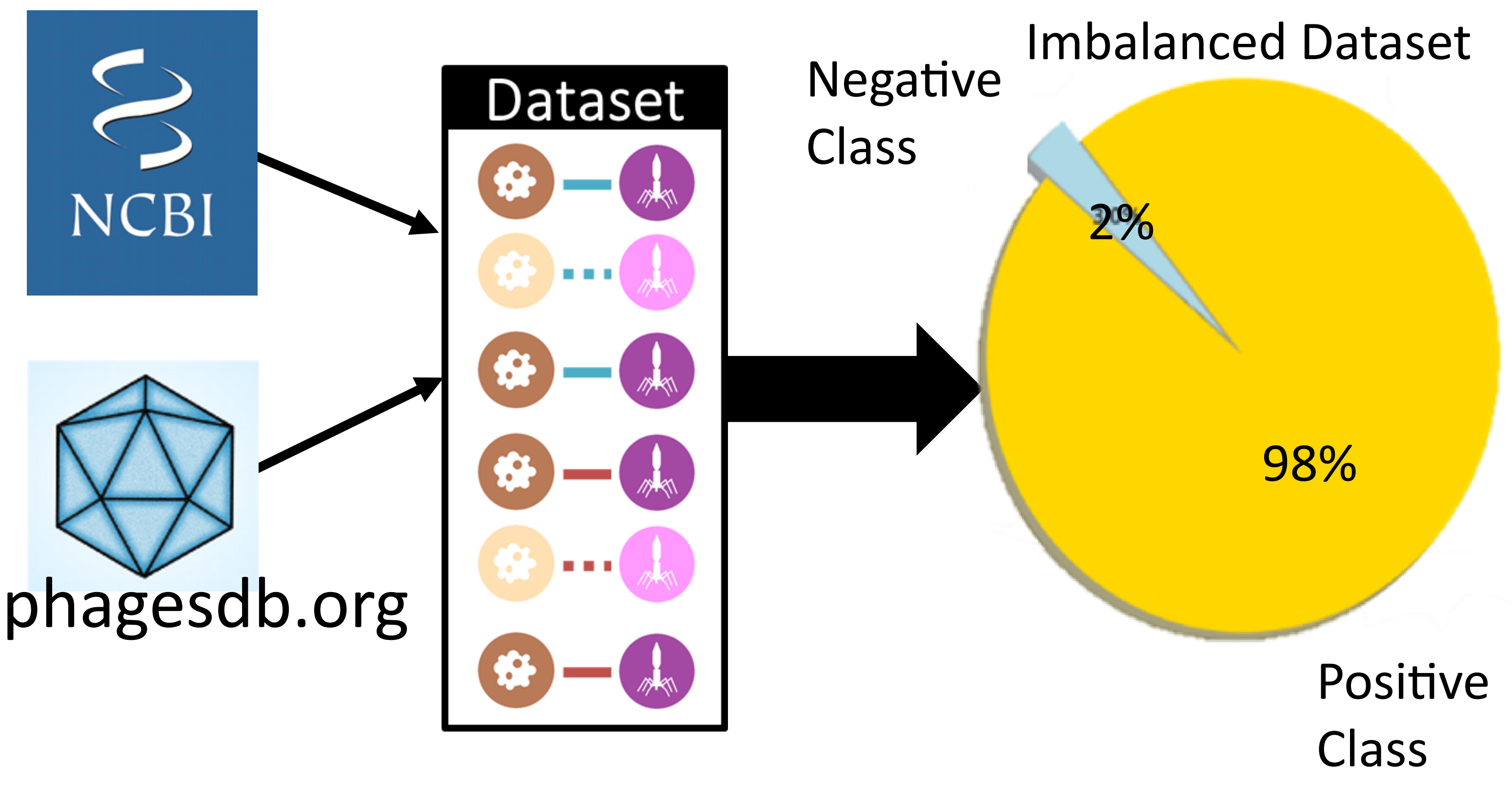


Elliptic Envelope (ELLEnv)

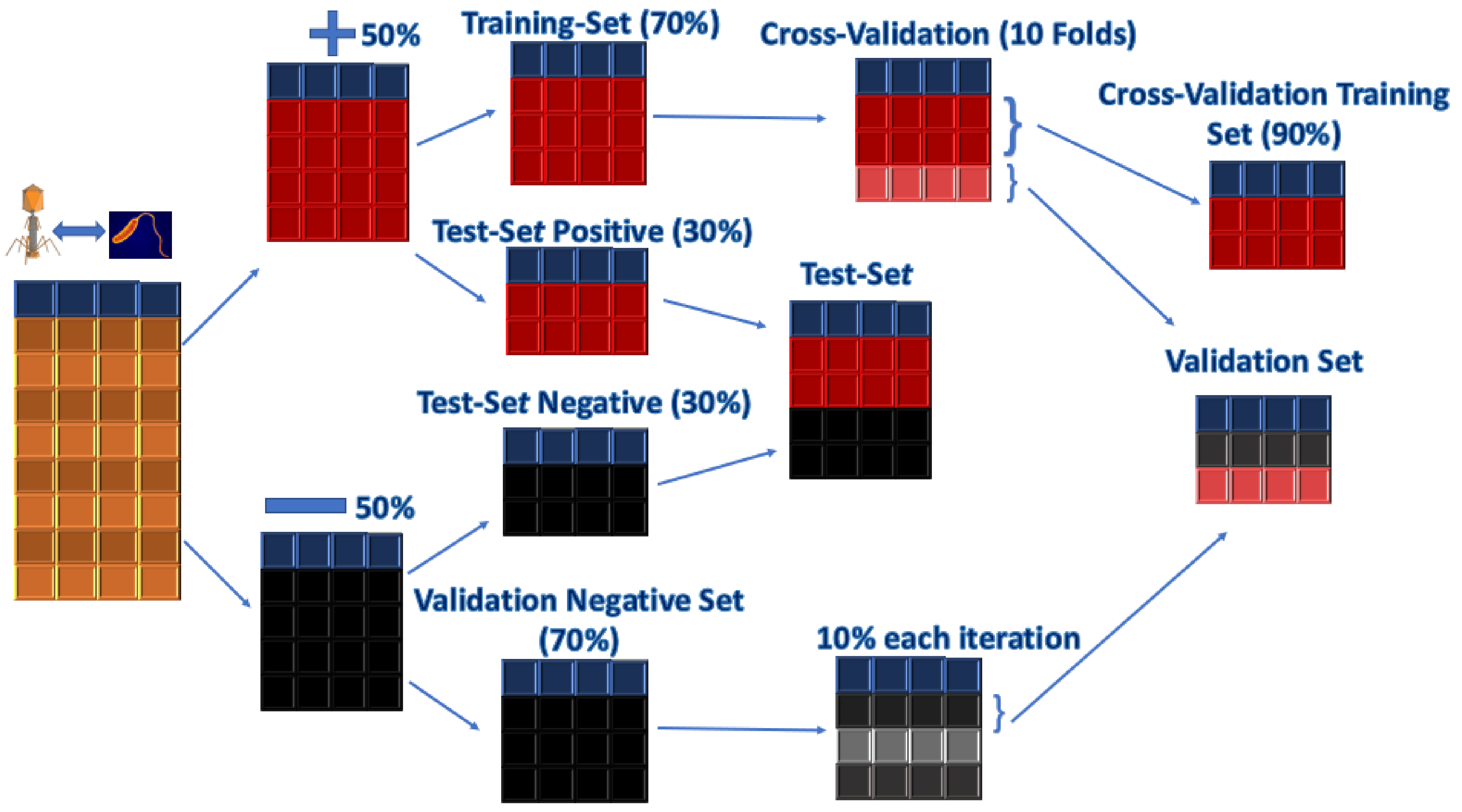



Dataset
Replicator Neural Network (RNN)
Elliptic Envelope (ELLEnv)
Local Outlier Factor (LOF)
Isolation Forest (iForest)
One-Class Support Vector Machine (O-SVM)
Voting
Prediction

Background



Dataset Adaptation



Current Results

Individual Performance

Algorithm	Accuracy	F1-Score	Sensitivity	Specifity	PPV	NPV
ELLEnv	79.7%	80.7%	84.9%	74.4%	76.9%	83.1%
iForest	68%	73.3%	87.4%	48.5%	63%	79.3%
RNN	75.6%	78.1%	86.8%	64.2%	70.9%	82.9%
LOF	53.9%	67%	93.8%	14.1%	52.2%	69.3%
O-SVM	72.2%	73.7%	77.6%	66.9%	70.2%	74.8%

Ensemble Performance (5 voting)

Accuracy	F1-Score	Sensitivity	Specifity	PPV	NPV
72.7%	76.7%	89.6%	55.8%	67.1%	84.2%

Ensemble Performance (3 best voting)

Accuracy	F1-Score	Sensitivity	Specifity	PPV	NPV
82.3%	83.2%	87.5%	77.1%	79.3%	86.1%

Conclusions

- The proposed approach seems very effective to deal with our imbalanced phage-bacteria interactions datasets. It exhibits accuracy and F1-score results above 80%.
- Some of the tested One-Class Learning algorithms tend to over-fit on the positive data, while others are more robust exhibiting good results in every metric, including those related with negative outcomes. The most robust method appears to be Elliptic Envelope.
- Applying ensemble-learning, with the top-three algorithms, improves the general performance for every used metric. This implies that one-class learning may benefit of such synergistic approach.