Master Degree in Computer Engineering - Graduate Thesis

Academic Year 2017/2018

Development and Evaluation of an Attack Detection System in a Computer Network

Supervisor:
Prof. Francesco Marcelloni
Prof. Alessio Bechini

Candidate:
Stefano Cicero

Roadmap

- Introduction
- ANIDS based on GA and Fuzzy Logic
- UGR’16 Dataset
- ANIDS Improvements
- Performance Evaluation
- Conclusions

Introduction

Motivations

Computer networks changed the paradigm in which people perform some of their daily duties and operations:

- Home banking
- Voice over IP (VoIP)
- Internet of Things (IoT)
- E-commerce
- Video streaming
- …

Due to advancement in Internet technologies and the concomitant rise in the number of network attacks, network intrusion detection has become a significant research issue.

Intrusion or threat

Deliberate and unauthorized attempt to:

- access information
- manipulate information
- render a system unreliable or unusable

IDSs - Intrusion Detection Systems

- Monitor and analyze user, system and network activities
- Configure systems for generation of reports of possible vulnerabilities
- Assess system and file integrity
- Recognize patterns of typical attacks
- Analyze abnormal activity
- Track user policy violation

IDSs classification

Deployment:

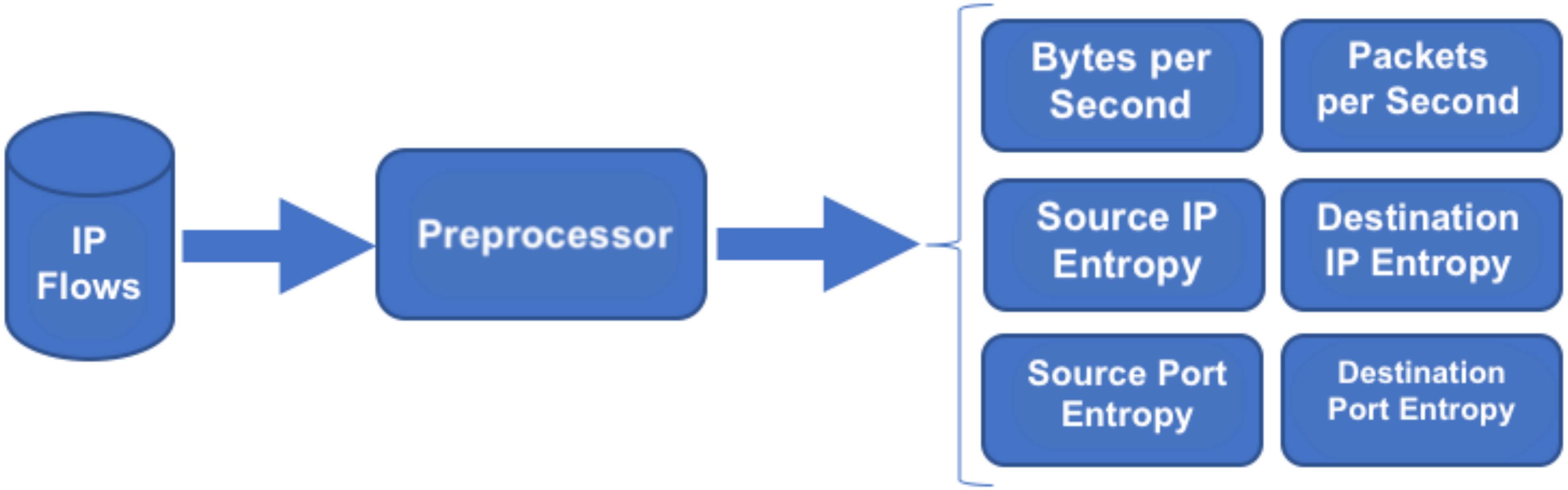
- host-based IDS (HIDS)
- network-based IDS (NIDS)

Detection mechanism:

- misuse (signature)-based
- anomaly-based
- hybrid

ANIDS based on GA and Fuzzy Logic

Data preprocessing



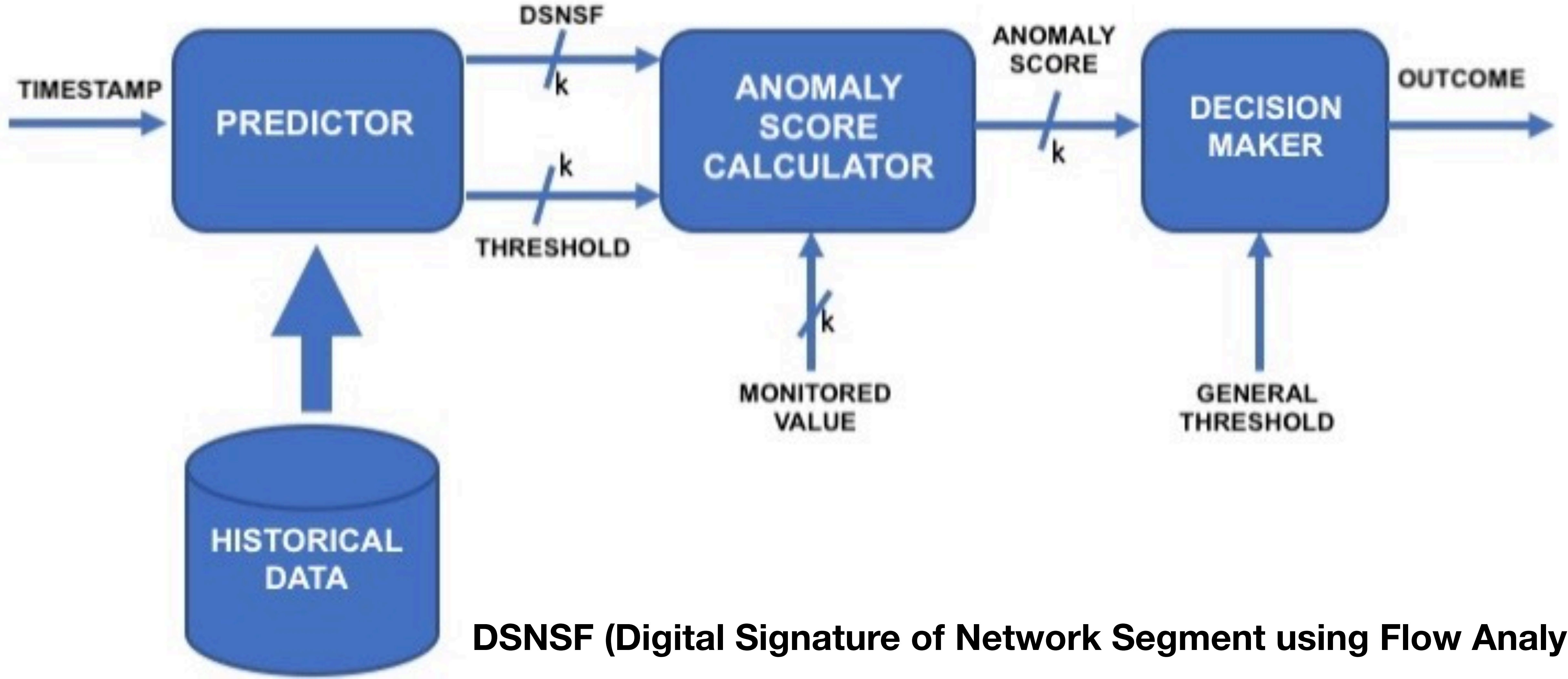
Shannon Entropy

$$H(X) = -\sum_{i=1}^{s} \left(\frac{x_i}{\sum_{i=1}^{s} x_i}\right) \log_2 \left(\frac{x_i}{\sum_{i=1}^{s} x_i}\right)$$

$X = attribute = \{x_1, \ldots, x_i, \ldots, x_s\}$

$x_i =$ Frequency of occurrence of sample *i* in the time interval

System Architecture

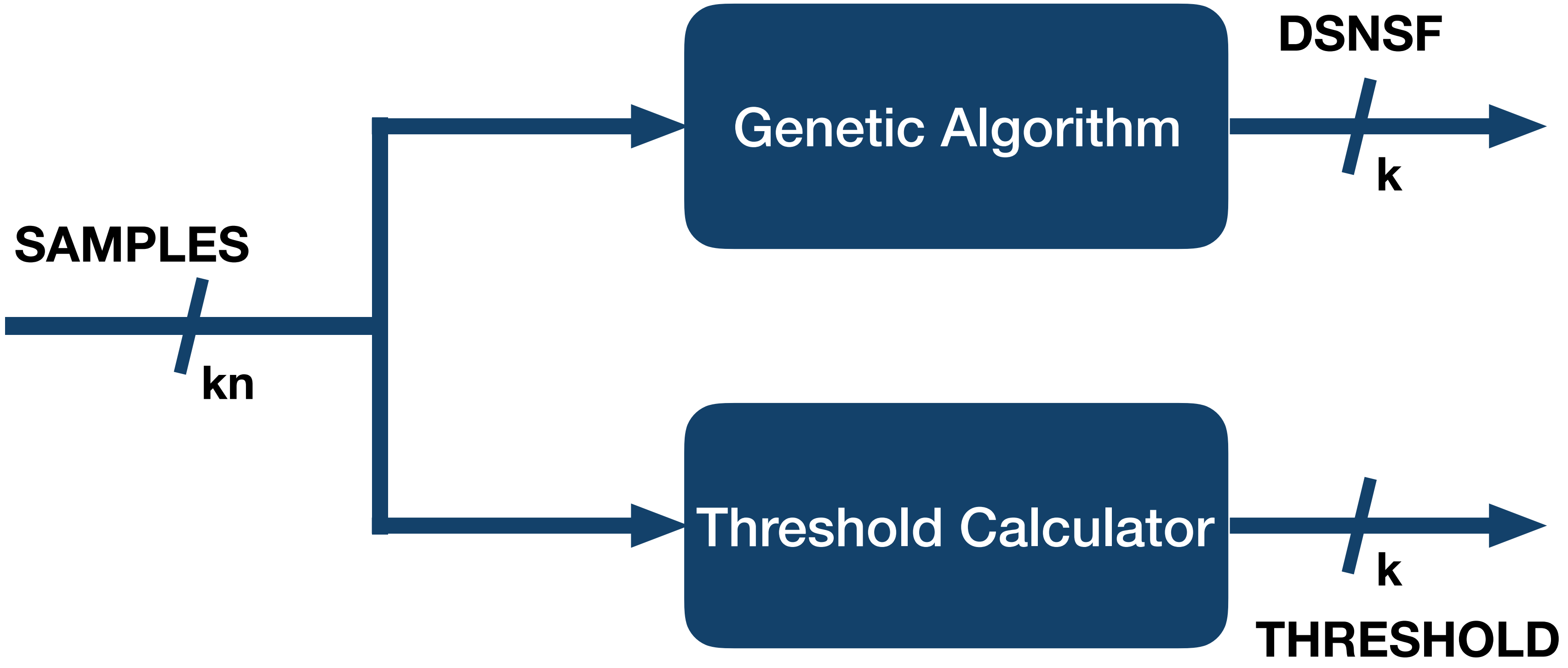


DSNSF (Digital Signature of Network Segment using Flow Analysis)

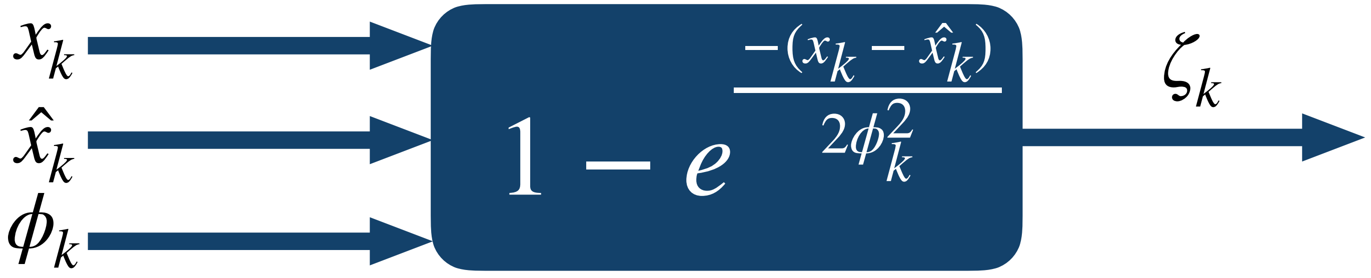
k = number of features

TIMESTAMP = day of the week, hour, minute

Predictor


DSNSF
Genetic Algorithm
k
SAMPLES
kn
Threshold Calculator
k
THRESHOLD

Anomaly Score Calculator



x_k

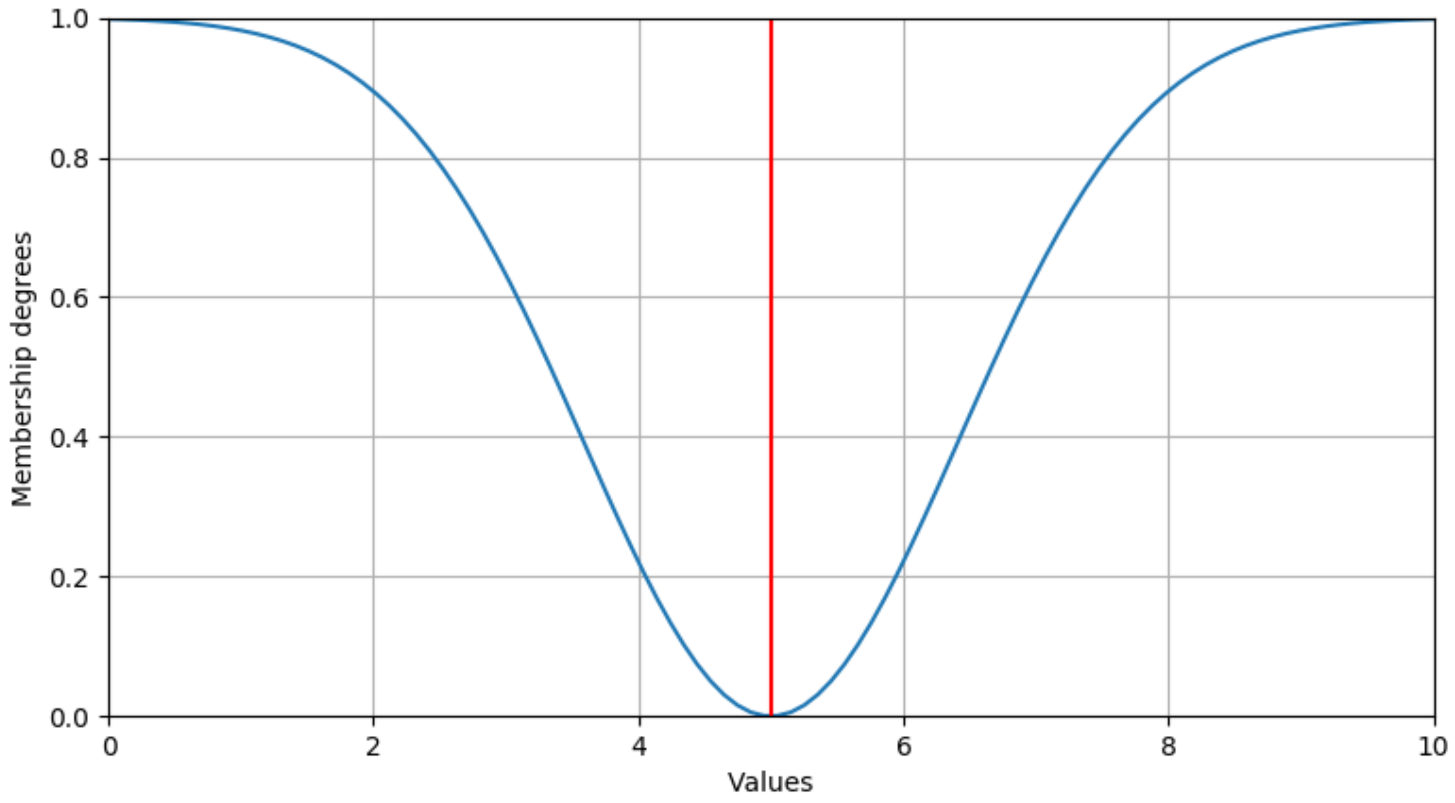
$\hat{x}_k$

ϕ_k

$$1 - e^{\frac{-(x_k - \hat{x}_k)}{2\phi_k^2}}$$

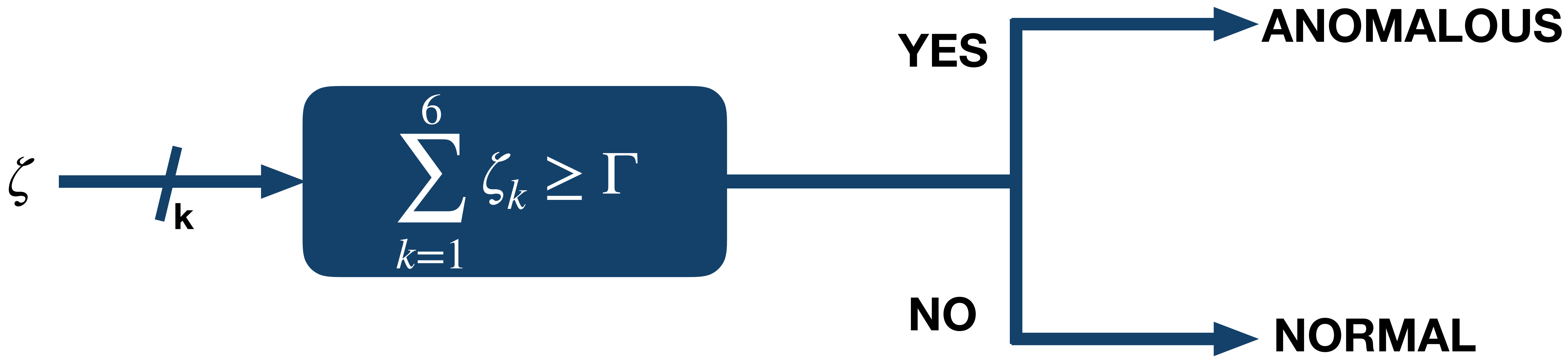
ζ_k

Example $\hat{x}_k = 5, \phi_k = 2$



Symbol	Meaning
x_k	Monitored value
$\hat{x}_k$	Predicted value (DSNSF)
ϕ_k	Threshold
ζ_k	Anomaly score

Decision Maker

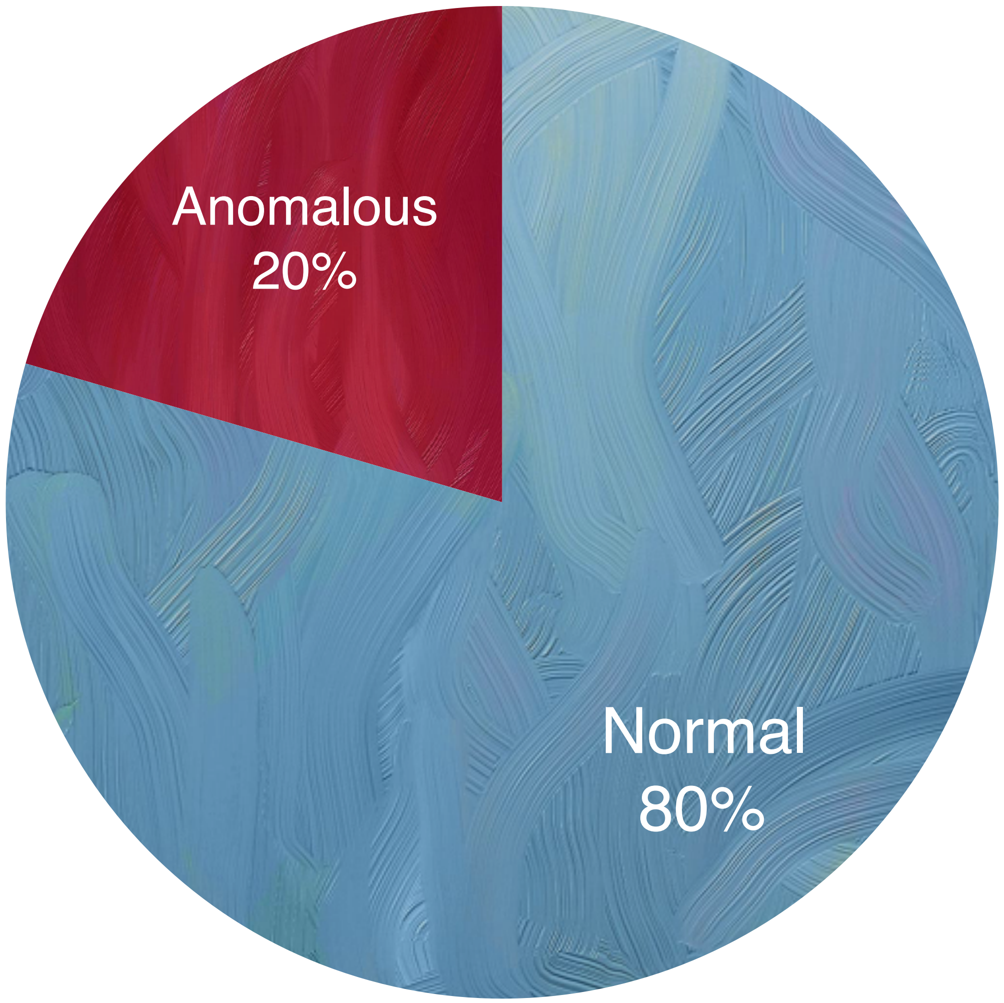


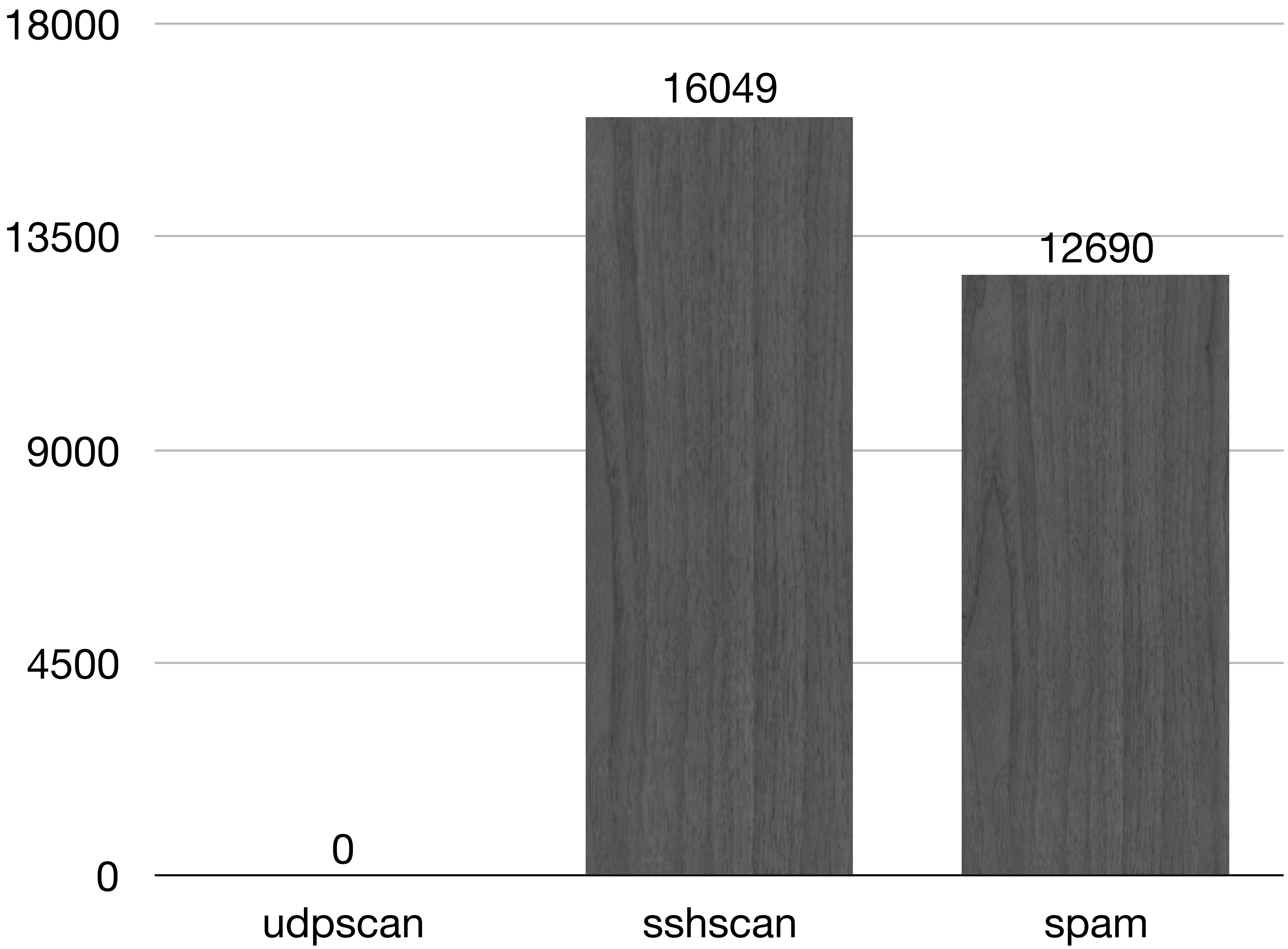
UGR’16 Dataset

UGR’16: Dataset Capture

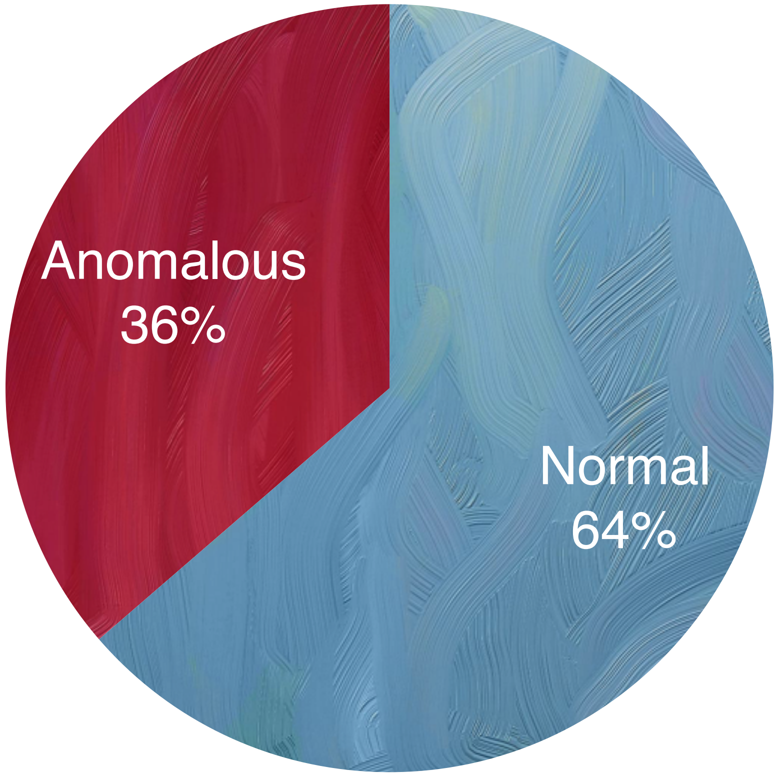
Feature	Calibration	Test
Capture start	10:47h 03/18/2016	13:38h 07/27/2016
Capture end	18:27h 06/26/2016	09:27h 08/29/2016
Attacks start	N/A	00:00h 07/28/2016
Attacks end	N/A	12:00h 08/09/2016
Number of files	17	6
Size (compressed)	181GB	55GB
# Connection	≈ 13000M	≈ 3900M

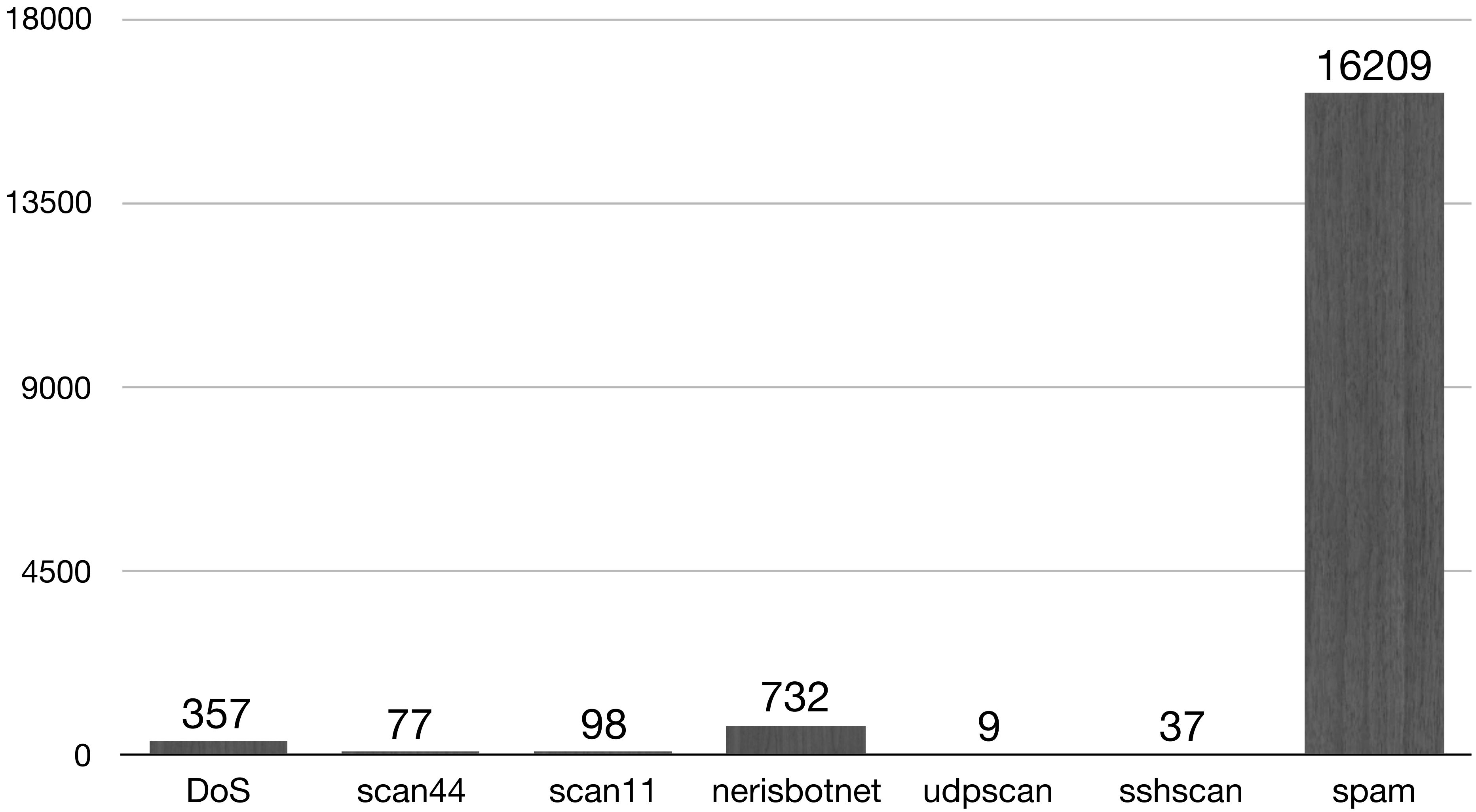
Training Set Composition





Test Set Composition





ANIDS Improvements

Changes and Check List

- Replaced genetic algorithm with mean value
- Removed features “byte per second” and “packet per second”
- Checked effectiveness of entropy features
- Added features “flag entropy” and “number of SMTP flows”
- Checked features values distribution

Performance Evaluation

Scenarios

Name	Training Set	Test Set
Scenario 1	Original training set	Original test set
Scenario 2	Original training set	Test set without “anomaly-sshscan”, “anomaly-udpscan”, “anomaly-spam” flows
Scenario 3	Original training set	Original training set
Scenario 4	Training set without anomalous flows	Original test set

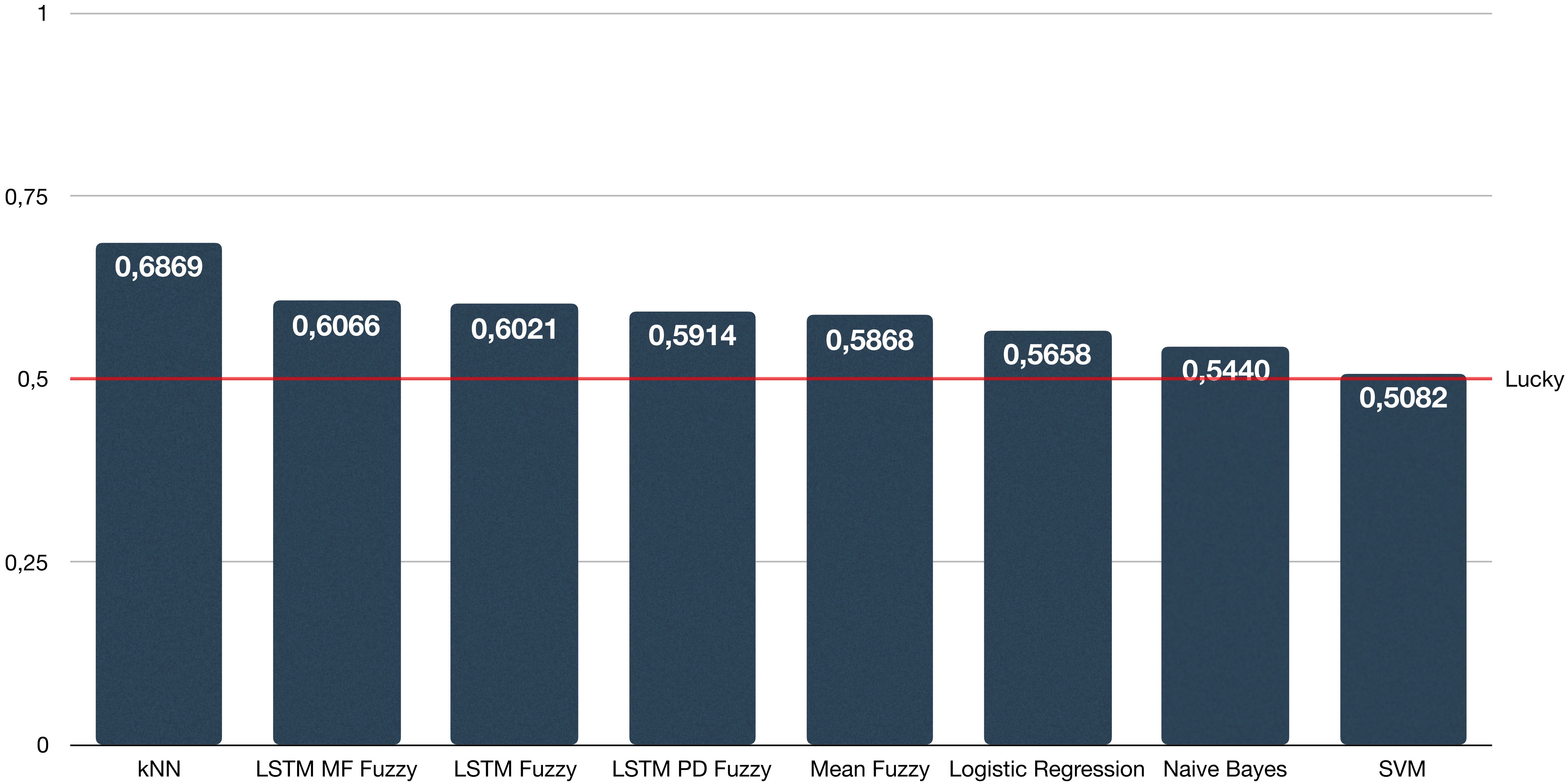
Metrics

- Confusion matrix
- Receiver Operating Characteristics (ROC) curve
- ROC Area Under the Curve (AUC)
- Execution time

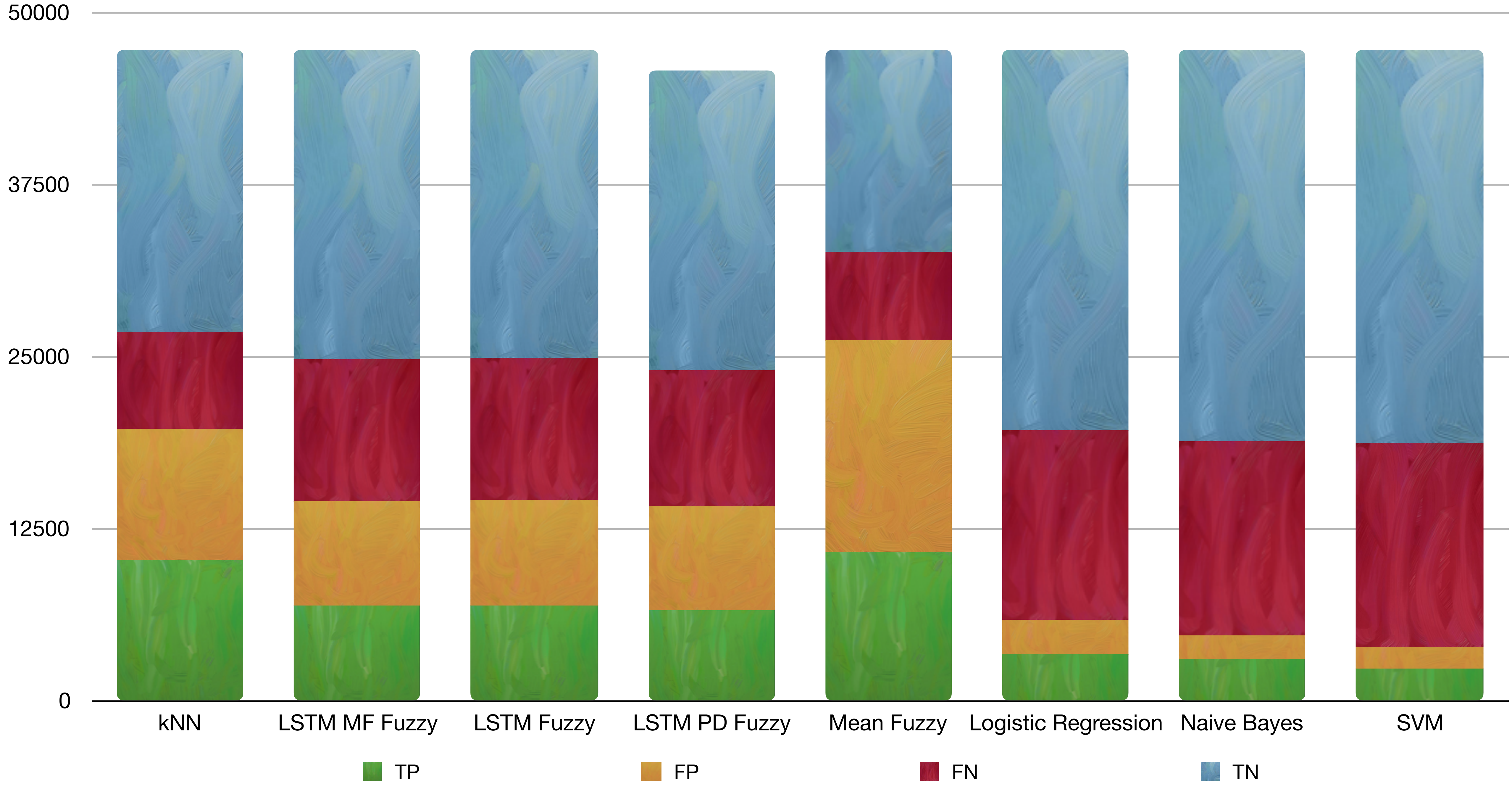
Comparison Systems

- kNN (k Nearest Neighbor)
- SVM (Support Vector Machine)
- Naive Bayes
- Logistic Regression
- LSTM Systems

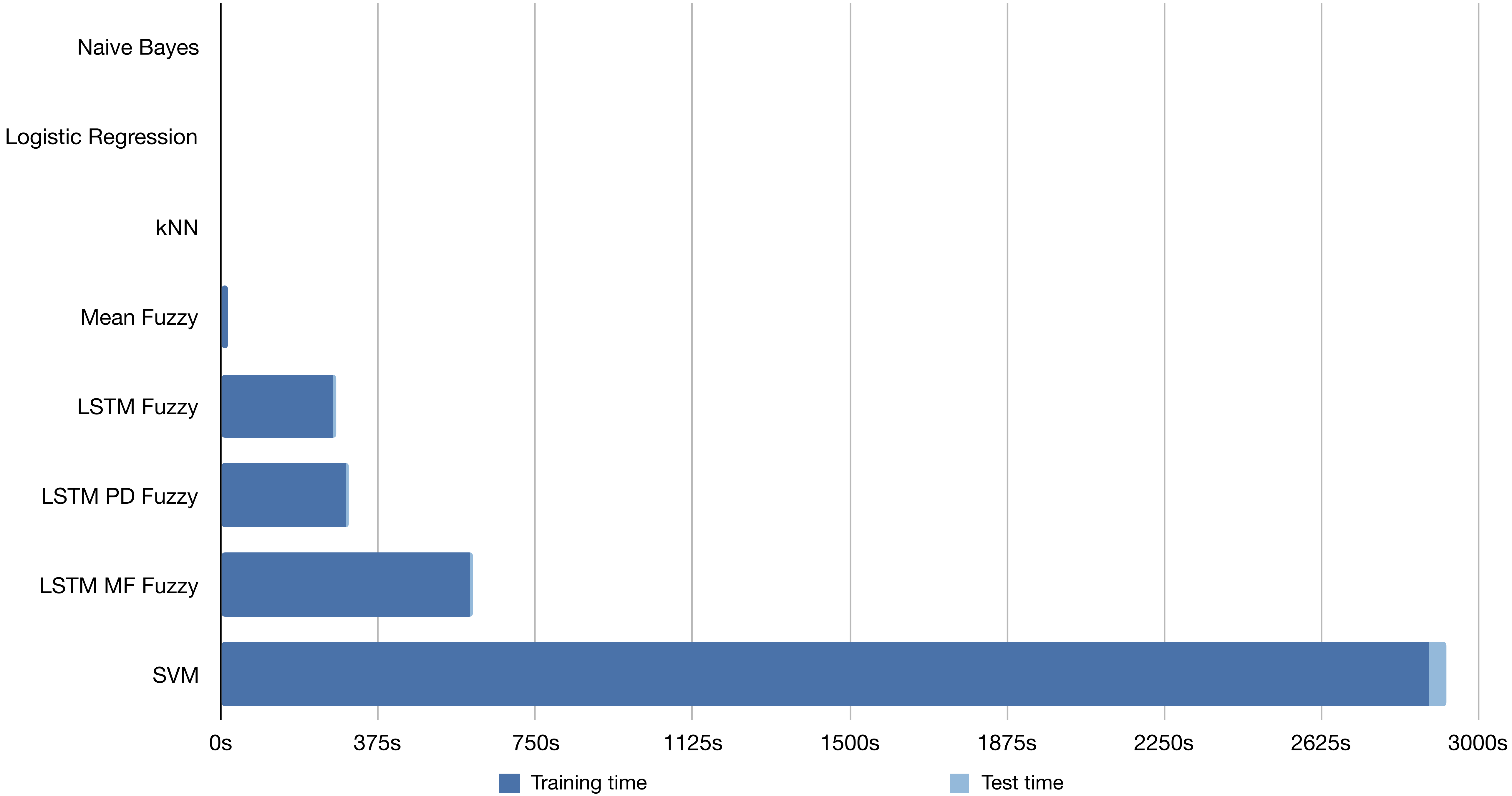
AUC Score - Scenario 1



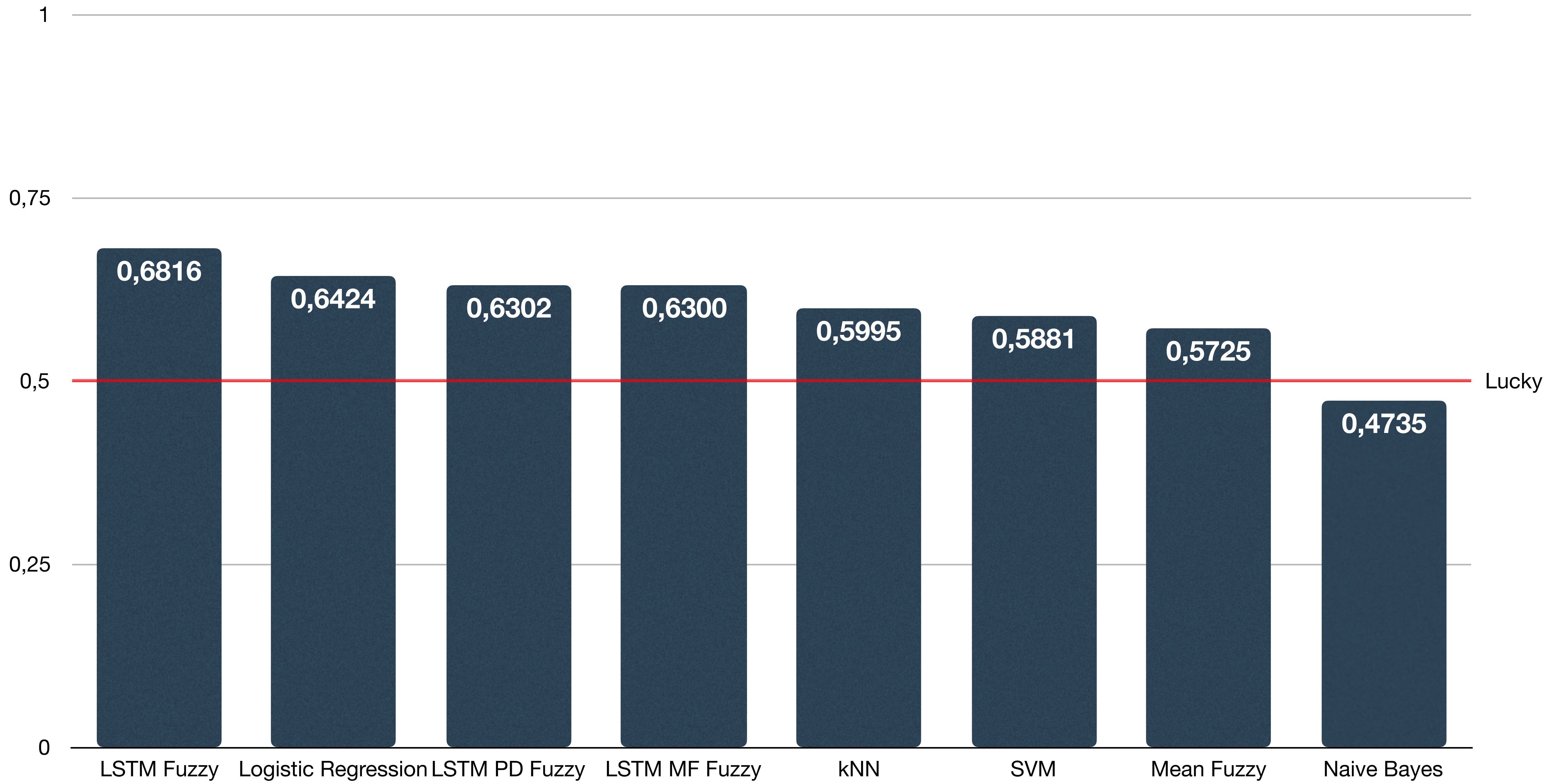
Confusion Matrix - Scenario 1



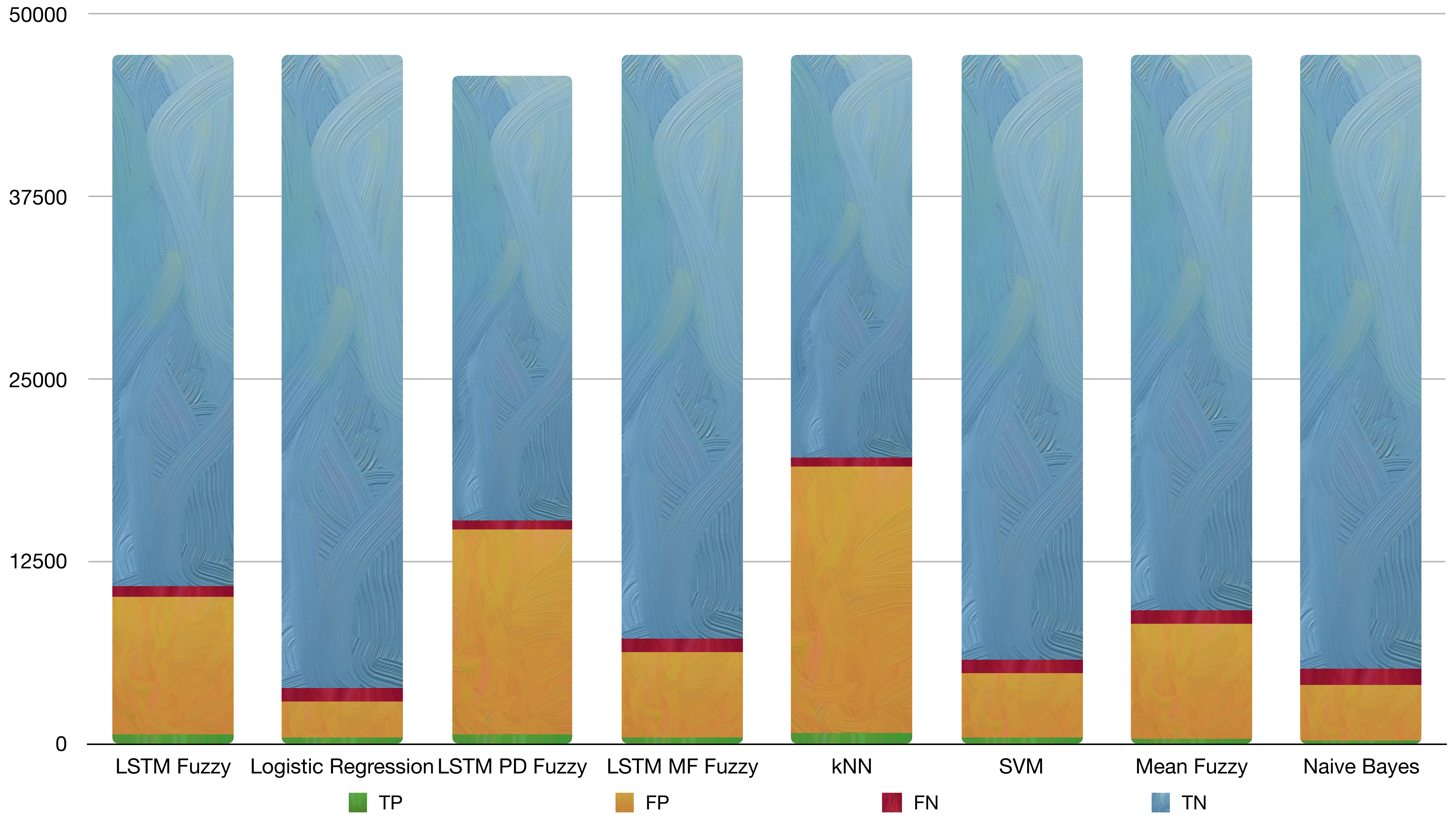
Execution Time - Scenario 1



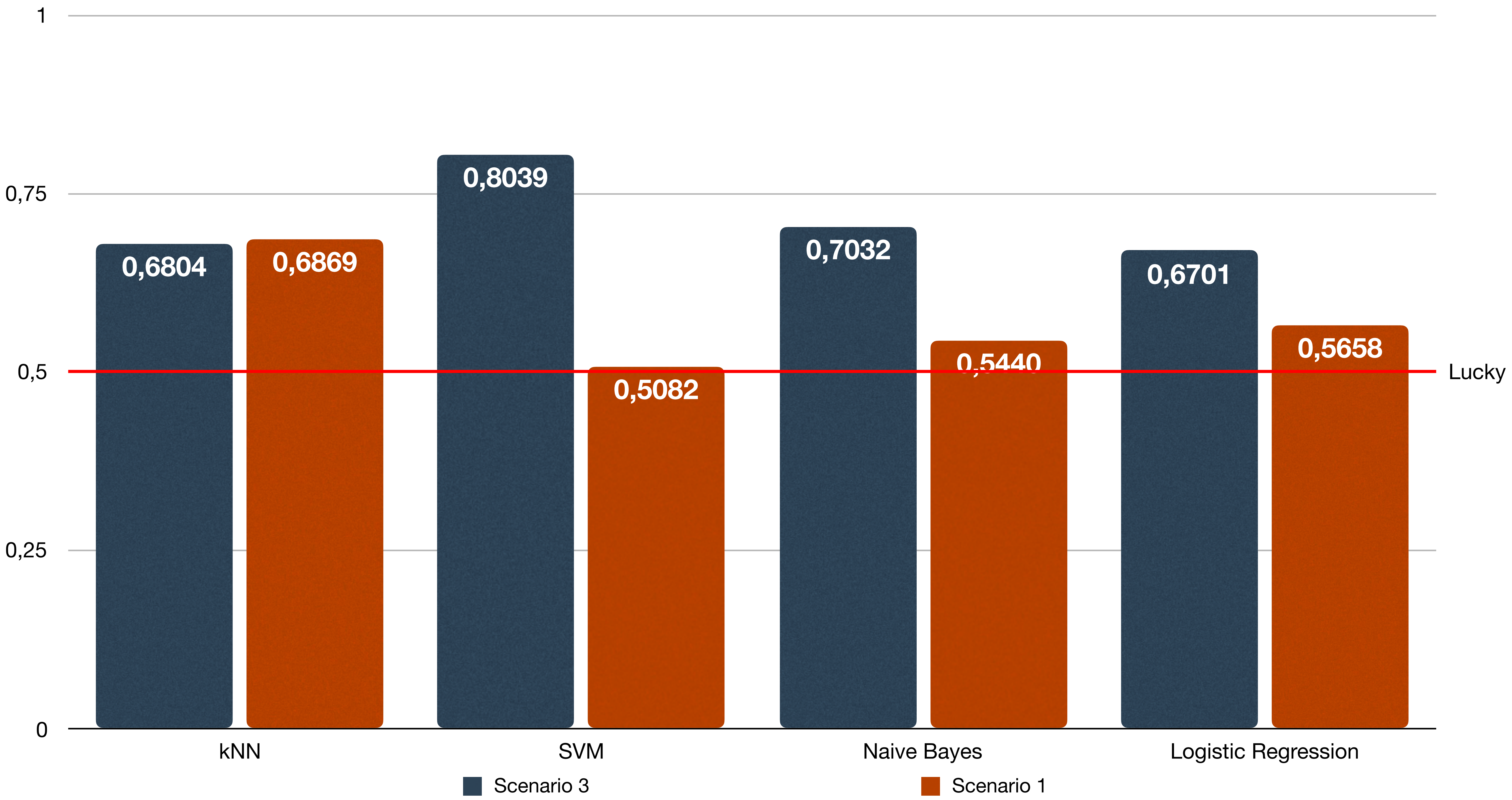
AUC Score - Scenario 2



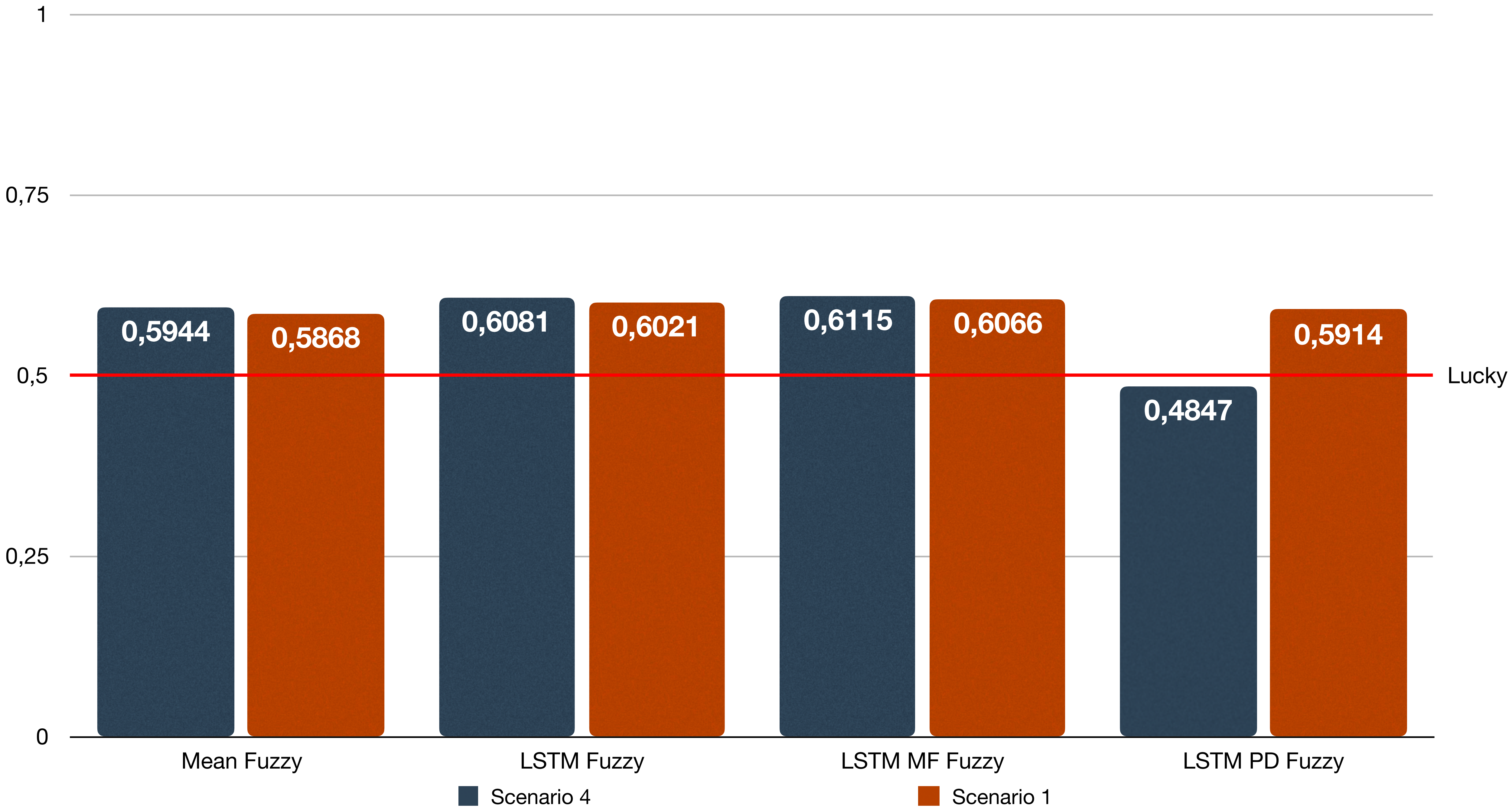
Confusion Matrix - Scenario 2



AUC Score - Scenario 3 vs Scenario 1



AUC Score - Scenario 4 vs Scenario 1



Conclusions

- Analysis, development, improvement and evaluation of a new ANIDS based on soft computing techniques

- Analysis and preprocessing of a new dataset for ANIDS evaluation

- kNN reached the best AUC score in Scenario 1, and it is one of the fastest system evaluated

- In Scenario 2 LSTM Fuzzy has the highest AUC Score but Logistic Regression has a better confusion matrix and execution time

- Mean Fuzzy obtained similar results in different scenarios

Thank you