Academy of the Federal Guard Service of the Russian Federation

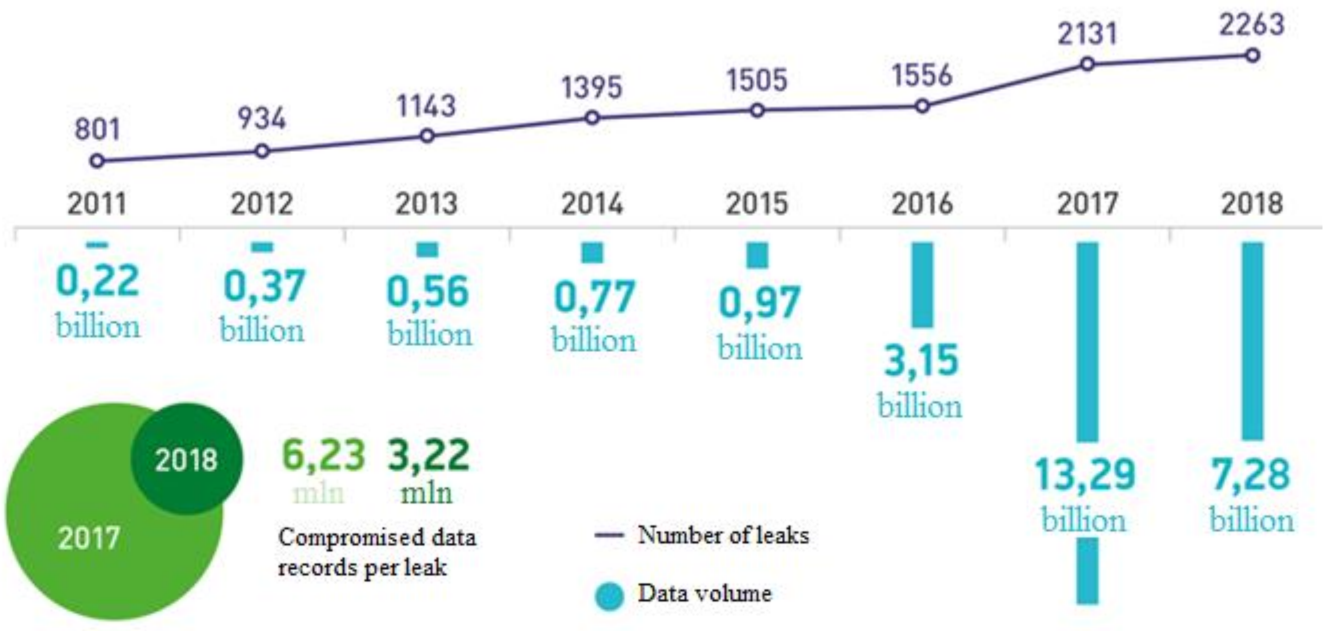
Classification of pseudo-random sequences based on the random forest algorithm

Speaker: Andrey A. Spirin , employee

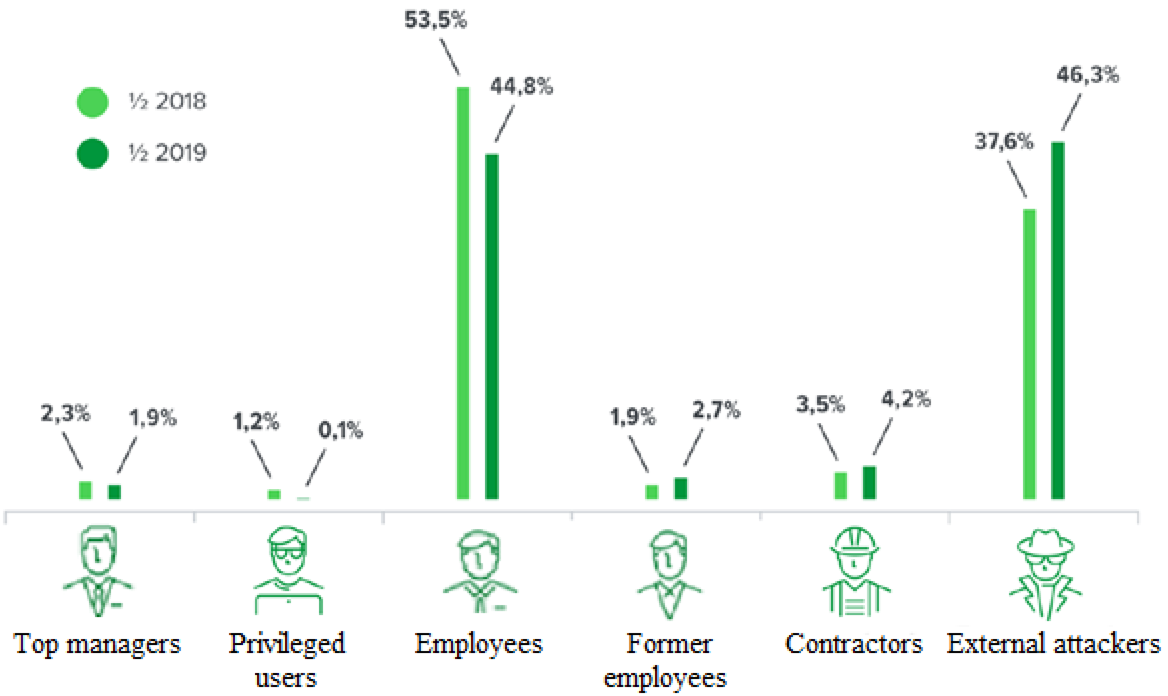
Scientific advisor: Alexander V. Kozachok, Dr. Sci., employee

25 september 2020

Analysis of the research object



a)



b)

Fig. 1. Recorded information leaks statistics *:
a) by the number of leaks and the number of leaked records,
b) by source (the person responsible).

* According to the InfoWatch analytical center

Research subject analysis I

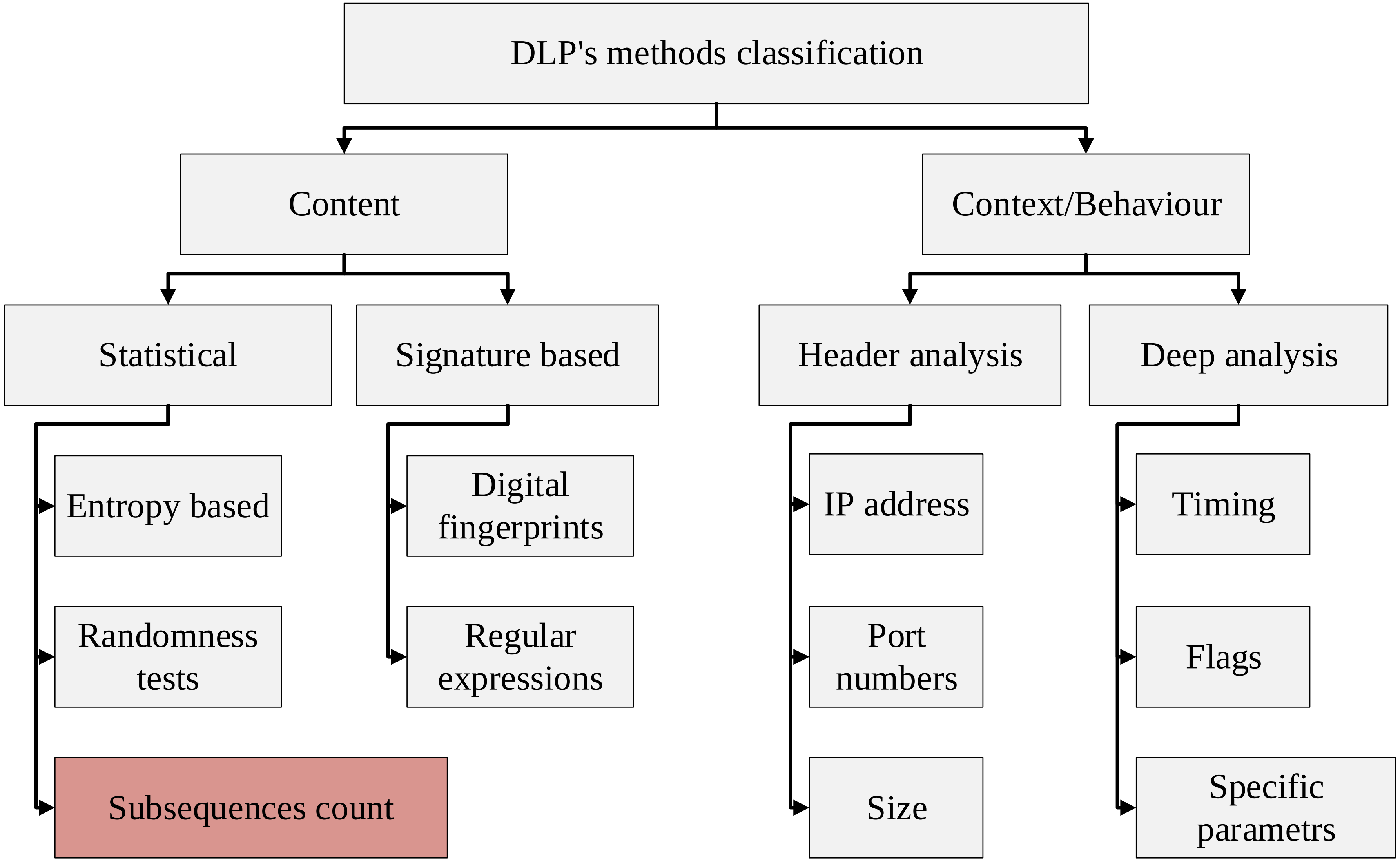


Fig. 2. DLP's methods classification

Research subject analysis II

Tab. 1. Overview of research in the subject area

Authors	Year	Task	Features	Algorithm	Results
H. Zhang, C. Papadopoulos, D. Massey	2013	Encrypted botnet traffic identification	Flows and packets entropy	Monte-Carlo method	Probability: flow (SSH,HTTPS) -95% packet (SSH,HTTPS) - 97%
A. R. Khakpour, A. X. Liu	2013	Binary, text and encrypted data identification	Subsequences length [1-10] byte entropy	CART, SVM	Accuracy: 0.87
D. Hahn, N. Apthorpe, N. Feamster	2018	Encrypted and compressed data classification	Entropy, x-square	k - NN, CNN, FFN	Probability: k-NN 60.0% FFN 54.1% CNN 66.9%
F. Casino, K.-K. R. Choo, C. Patsakis	2019	Encrypted and compressed traffic classification	Absolut value and confidence interval of x-square test. NIST test: frequency block test, cumulative sum test, approximation entropy test.	HEDGE (High Entropy DistinGuishEr)	Accuracy: 0.71
Z. Tang, X. Zeng, Y. Sheng	2019	Encrypted and compressed traffic classification	Entropy of 4,8,16,24 bit subsequences	SVM, RF	Accuracy: for traffic 0.979 video 0.70 images, text 0.72 audio 0.66

Source data

Source data	Algorithm	Class label	Batch, files	File length, KB
Text	AES(CBC)	0	2000	600
	3DES(CBC)	0	2000	600
	Camellia(CBC)	0	2000	600
	RC4(CBC)	0	2000	600
	GOST 34.12 (ECB)	0	2000	600
Text	ZIP	1	2000	600
	RAR	1	2000	600
	7Z	1	2000	600
	XZ	1	2000	600
	GZ	1	2000	600
	BZ2	1	2000	600

Tab. 2. Source data for experiments

Research objective

$$F : X \to Y = \{0,1\} \qquad (1)$$

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

$$Recall(TPR) = \frac{TP}{TP + FN} \qquad (3)$$

$$FPR = \frac{FP}{FP + TN} \qquad (5)$$

$$AUC - ROC = F(TPR, FPR) \qquad (6)$$

where TP – number of objects correctly assigned to the class i, TN – number of objects correctly assigned to the class j (j≠i), FP – number of false positives (error type I), FN – number of target passes (error type II)

Stage I-I. NIST tests p-value features

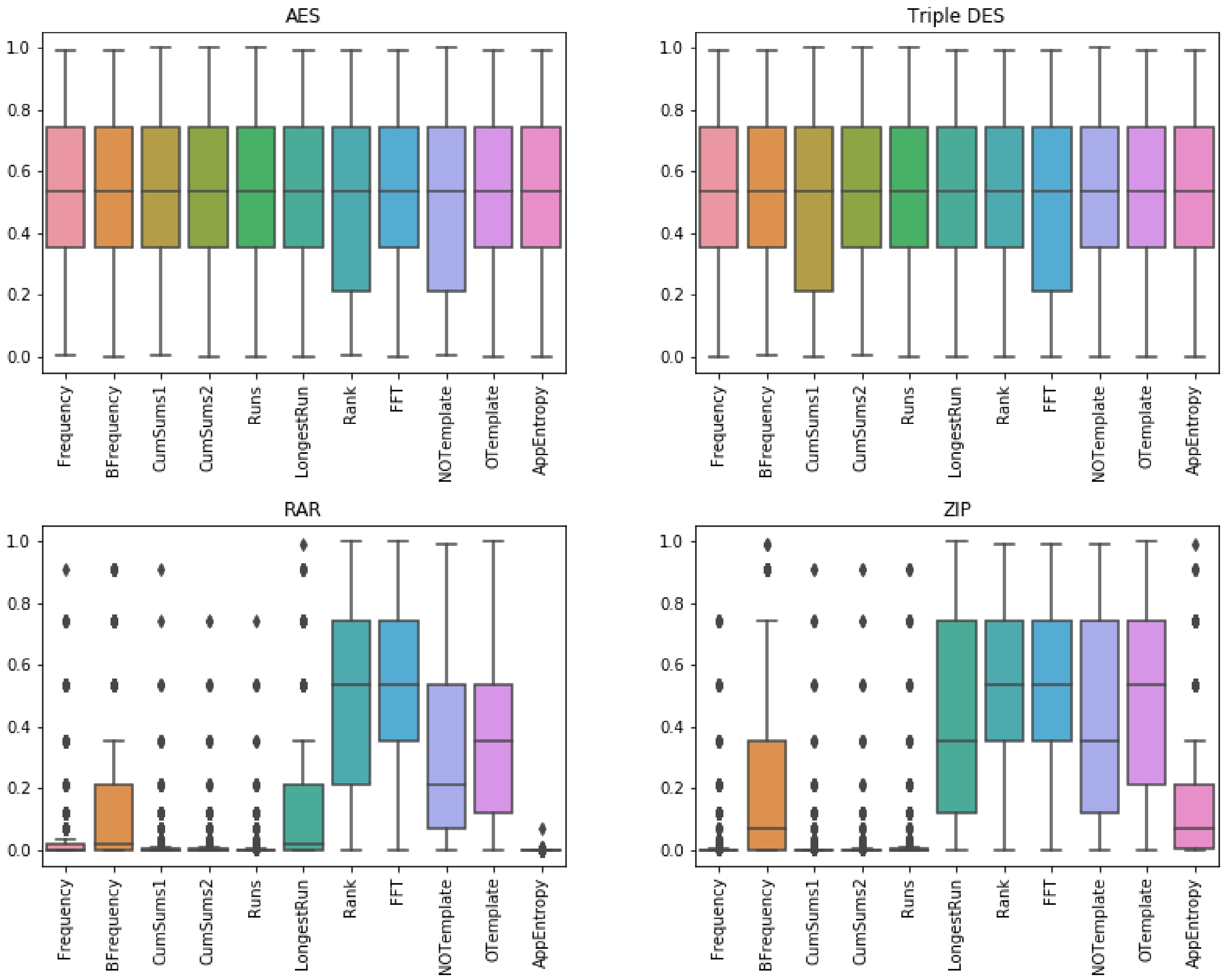


Fig. 3. Feature space based on passed NIST tests

Stage I-II. Machine learning algorithm selecting

Algorithm	Accuracy
Decision Tree	0,58
Random Forest	0,64

Tab. 3. Machine learning algorithms classification accuracy using the feature space formed by the passed NIST tests on a sample of compressed and encrypted sequences:

1 class – archives with extensions RAR, ZIP,
2 class – encrypted by AES, 3DES.

Stage II-I. Feature extraction

Data: P: |P|=Q, S: $|S| = 2^N$, B:$|B| = 256$
Result: $F_{Q,E}$

$F_{Q,E} \leftarrow <>$
for $p \in P$ **do**
  $M_p \leftarrow$ **Len**(p)
  for $s \in S$ **do**
    $n_s \leftarrow$ **Count**(p,s)
    $f_{p,s} \leftarrow \dfrac{n_s}{M_p - N_s + 1}$
    $F_{Q,E} \leftarrow F_{Q,E} \cup < s, f_{p,s}, y_i >$
  for $b \in B$ **do**
    $n_b \leftarrow$ **Count**(b,s)
    $bytes_p \leftarrow < b, n_b >$
  $F_{Q,E} \leftarrow F_{Q,E} \cup bytes_p$
  $std_p =$ **Std**$(bytes_p)$
  $min_p =$ **Min**$(bytes_p)$
  $max_p =$ **Max**$(bytes_p)$
  $delta_p = max_b - min_b$
  $F_{Q,E} \leftarrow F_{Q,E} \cup < std_p, min_p, max_p, delta_p >$
return $F_{Q,E}$

Fig. 4. Feature extraction algorithm

Stage II-II. Machine learning algorithm selecting

Algorithm	Accuracy
Random Forest	0,94
Decision Tree	0,87
K-nearest neighbors	0,88
Gradient Boosting	0,89

Tab. 4. Machine learning algorithm selecting

Stage II-III. Accuracy dependence from numbers of features

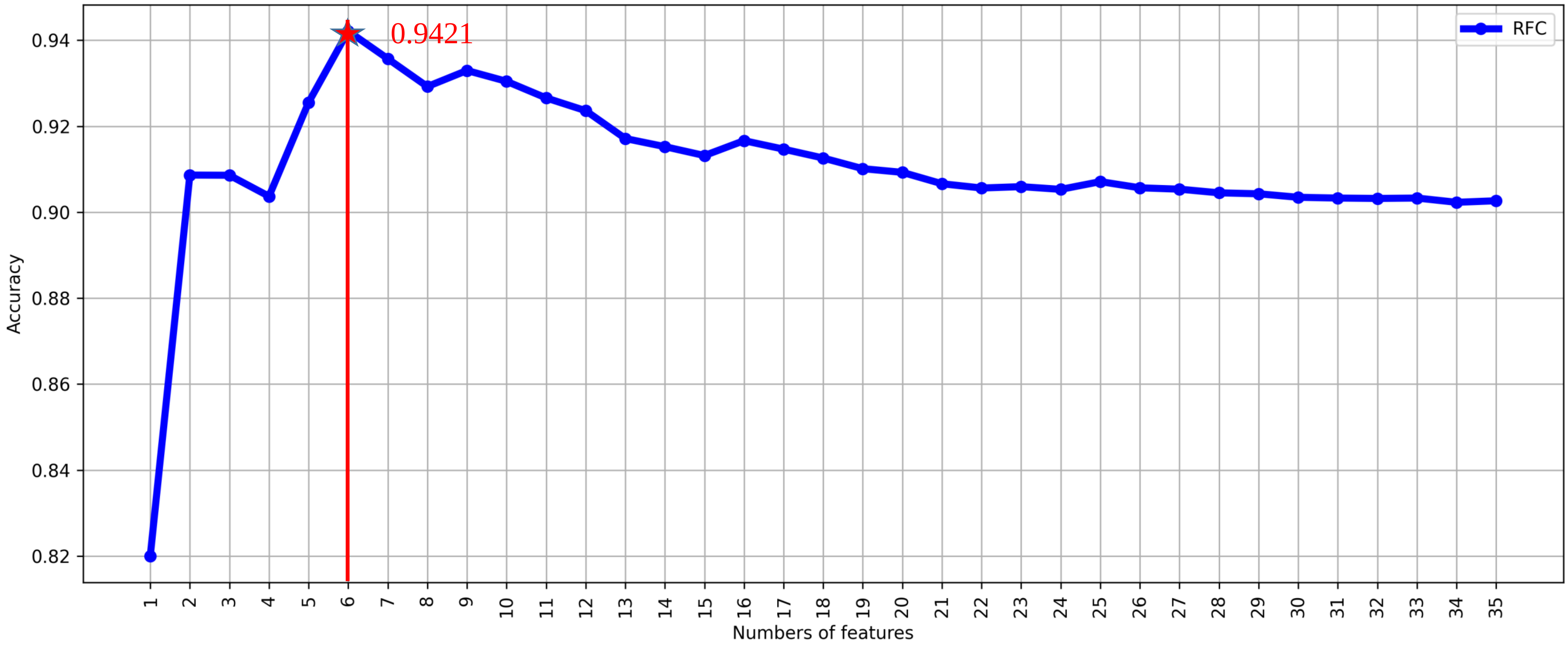


Fig. 5. Estimation of the dependence of classification accuracy on the number of features for the random forest algorithm

Stage II-IV. Accuracy dependence from tree depth

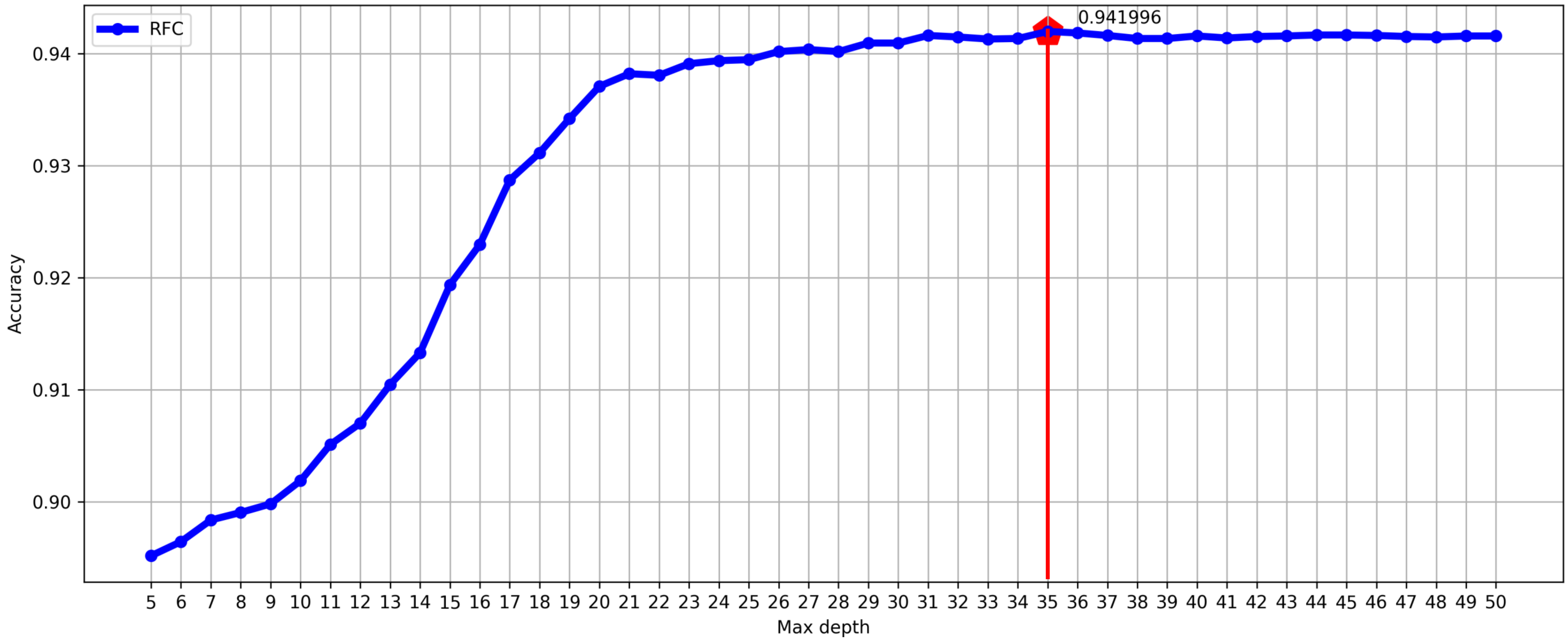


Fig. 6. Estimation of the classification accuracy dependence from the maximum tree depth of a random forest

Stage II-V. Accuracy dependence from number of trees

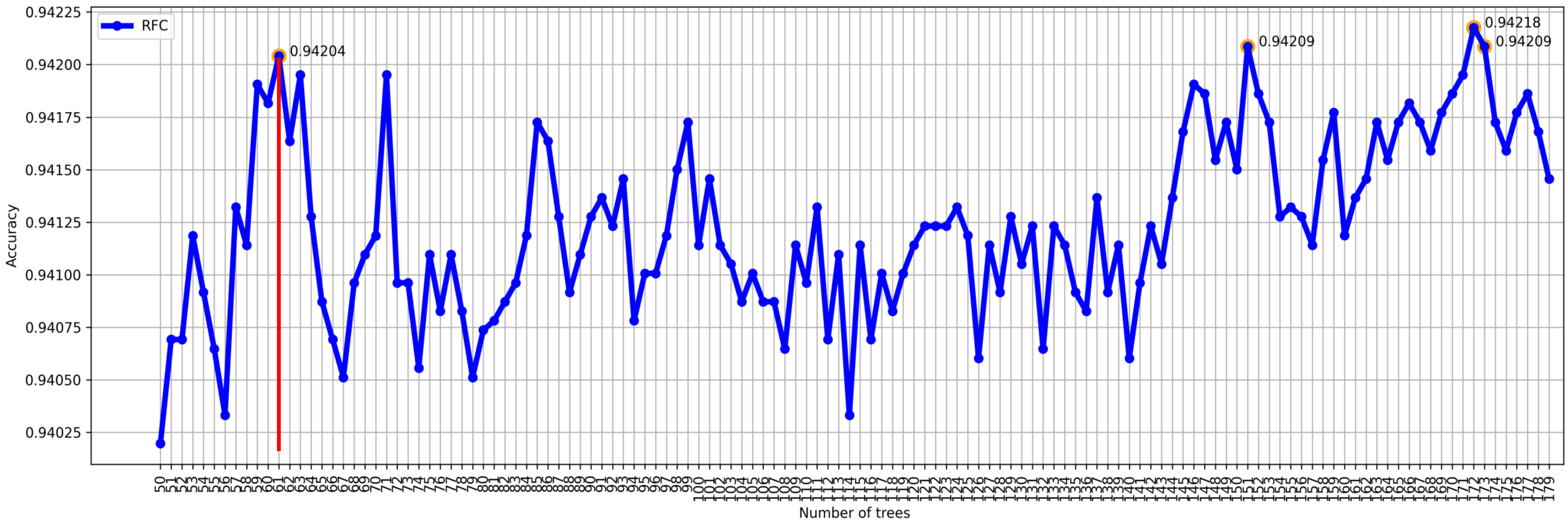


Fig. 7. Estimation of the classification accuracy dependence from the number of trees in a random forest

Stage II-VI. Accuracy dependence from the file size

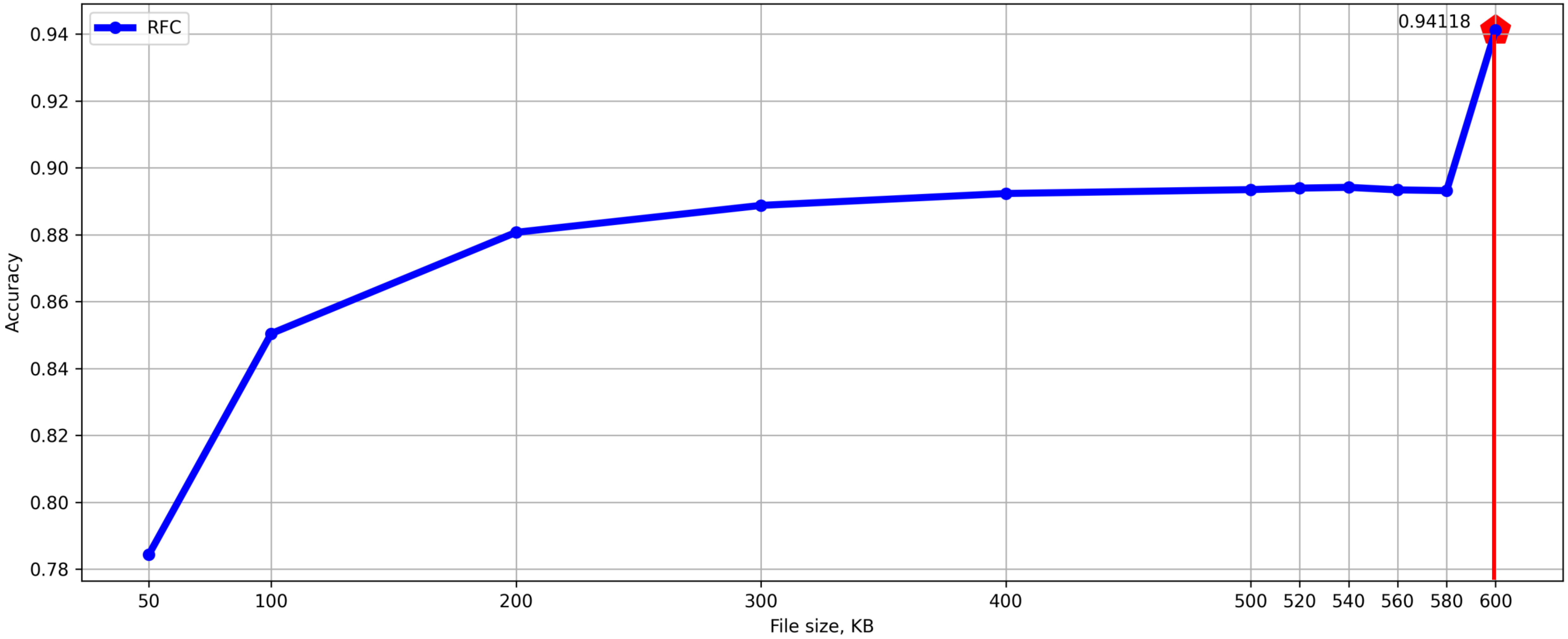


Fig. 8. Estimation of the classification accuracy dependence from the file size in KB

Stage II-VII. Classification accuracy dependence from type of compressed files

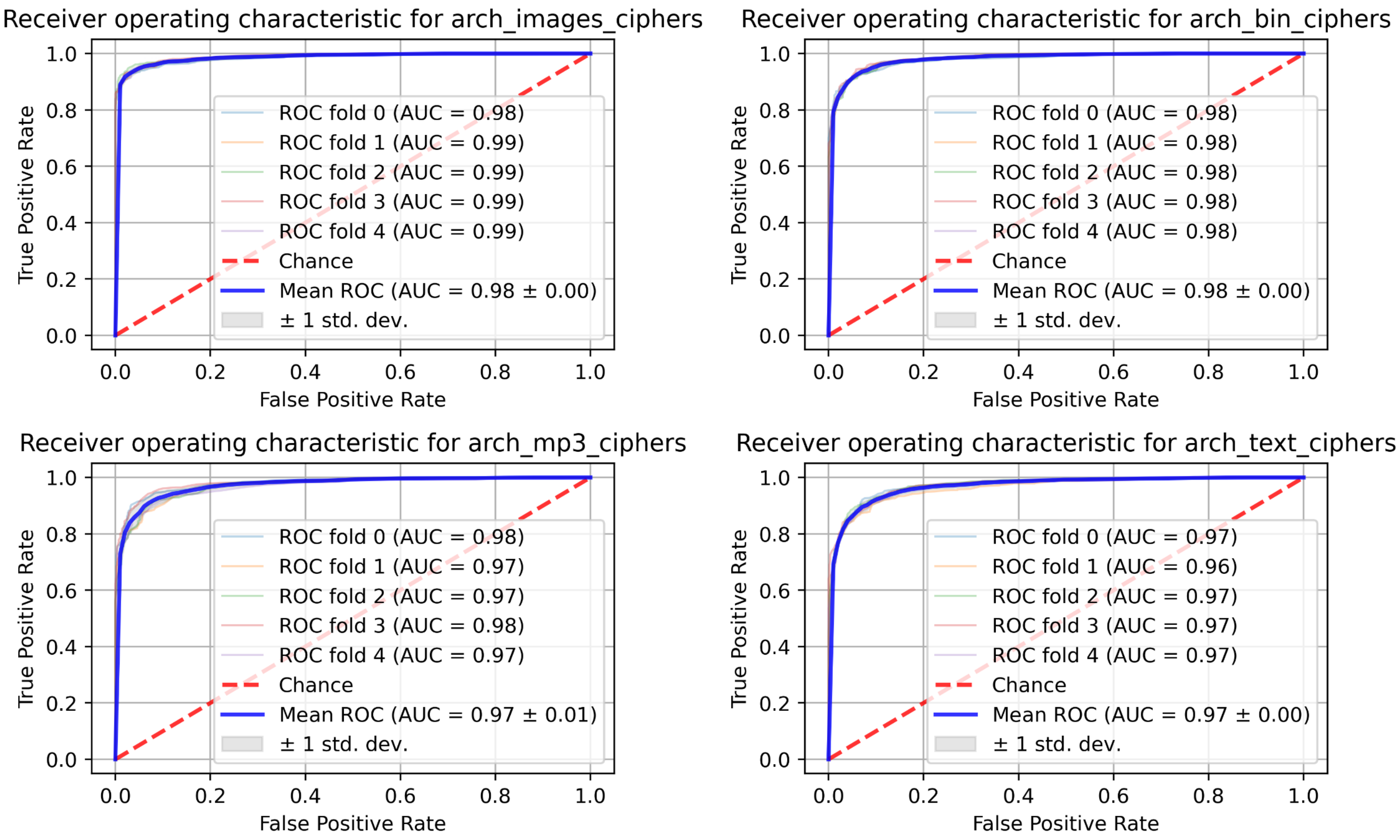


Fig. 10. Classification accuracy dependence from type of compressed files

Stage II-VIII. Results

Classificator hyper parameters	Value
Subsequences length	9 bits
Used features	6
Max tree depth	35
Number of trees	61

Class	Precision	Recall
Ciphers	0,92	0,98
Archives	0,98	0,93

Metrics	Value
AUC-ROC	0,99
F1-score	0,95
Accuracy	0,95

Features used by classifier
F(00000000)
max F(x)|x=0..255
F(100000000)
F(000000000)
F(000000001)
F(011111111)

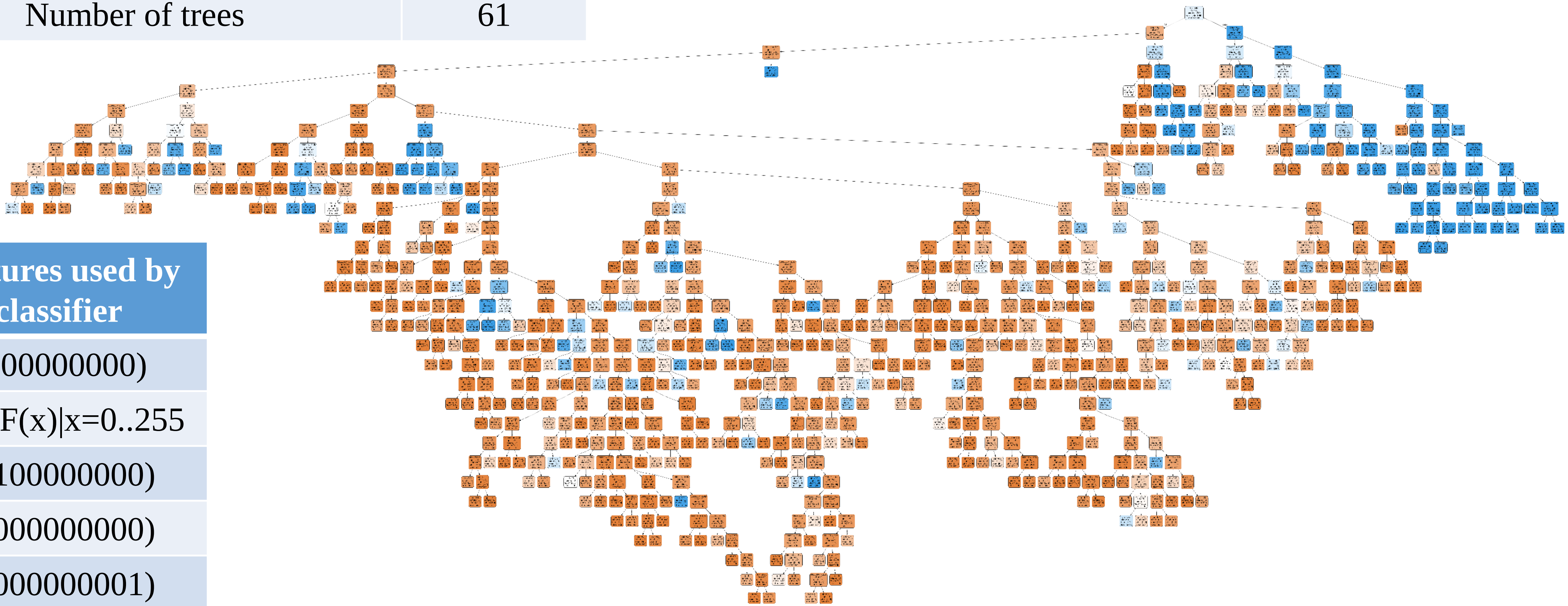


Fig. 11. Decision tree example in random forest classifier