

DOI: 10.15514/ISPRAS-2023-35(6)-11



## Анализ безопасности проекта национального стандарта «Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации»

*Г.Б. Маршалко, ORCID: 0009-0001-3499-3368 <marshalko\_gb@tc26.ru>*

*Р.А. Романенков, ORCID: 0009-0000-2670-4709 <romanenkov\_ra@tc26.ru>*

*Ю.А. Труфанова, ORCID: 0009-0003-5135-2196 <trufanova\_ua@tc26.ru>*

*Технический комитет по стандартизации «Криптографическая защита информации»,  
Россия, 127273, г. Москва, ул. Отрадная, 2Б стр.1.*

**Аннотация.** В работе предложен метод проверки принадлежности обучающей выборке для нейросетевого алгоритма классификации из проекта национального стандарта, разработанного Омским государственным техническим университетом под эгидой Технического комитета по стандартизации «Искусственный интеллект» (ТК 164). Указанный метод позволяет определить, использовались ли данные при обучении нейронной сети, и направлен на нарушение свойства конфиденциальности обучающей выборки. Полученные результаты показывают, что описываемый стандартом механизм защиты нейросетевых классификаторов не обеспечивает заявленных свойств. Полученные результаты первоначально анонсированы на конференции Рускрипто'2023.

**Ключевые слова:** Статистическая классификация; нейронные сети; информационная безопасность; обучающая выборка; атака проверки принадлежности обучающей выборке; конфиденциальность; стандартизация.

**Для цитирования:** Маршалко Г.Б., Романенков Р.А., Труфанова Ю.А., Анализ безопасности проекта национального стандарта «Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации». Труды ИСП РАН, том 35, вып. 6, 2023 г., стр. 179–188. DOI: 10.15514/ISPRAS–2023–35(6)–11.

## Security Analysis of the Draft National Standard «Neural Network Algorithms in Protected Execution. Automatic Training of Neural Network Models on Small Samples in Classification Tasks»

*G.B. Marshalko, ORCID: 0009-0001-3499-3368 <marshalko\_gb@tc26.ru>*

*R.A. Romanenkov, ORCID: 0009-0000-2670-4709 <romanenkov\_ra@tc26.ru>*

*J.A. Trufanova, ORCID: 0009-0003-5135-2196 <trufanova\_ua@tc26.ru>*

*Technical Committee for Standardization «Cryptographic Information Protection»,  
bld. 1, 2 B, Otradnaya st., Moscow, 127273, Russia.*

**Abstract.** We propose a membership inference attack against the neural classification algorithm from the draft national standard developed by the Omsk State Technical University under the auspices of the Technical Committee on Standardization «Artificial Intelligence» (TC 164). The attack allows us to determine whether

the data were used for neural network training, and aimed at violating the confidentiality property of the training set. The results show that the protection mechanism of neural network classifiers described by the draft national standard does not provide the declared properties. The results were previously announced at Ruscrypto'2023 conference.

**Keywords:** Statistic classification; neural networks; informational security; training set; membership inference attack; confidentiality; standardization.

**For citation:** Marshalko G.B., Romanenkov R.A., Trufanova J.A., Security analysis of the draft national standard «Neural network algorithms in protected execution. Automatic training of neural network models on small samples in classification tasks». *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 6, 2019. pp. 179-188 (in Russian). DOI: 10.15514/ISPRAS-2023-35(6)-11.

## 1. Введение

Работа посвящена исследованию безопасности проекта национального стандарта «Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации» [2] и [3], разработанного в 2022 году под эгидой Технического комитета по стандартизации «Искусственный интеллект» (ТК 164) Омским государственным техническим университетом.

Для технологий искусственного интеллекта, и, в частности для получивших в настоящее время наибольшее распространение технологий машинного обучения, как и для любой другой информационной технологии, необходимо рассматривать вопросы безопасности. Эти технологии обладают той особенностью, что алгоритм решения задачи (обученная модель), например, задачи классификации, получается в процессе ее решения и существенным образом зависит от входных данных (обучающей выборки), а не фиксирован заранее. Это создает новые способы построения атак на системы, использующие технологии искусственного интеллекта, не актуальные для других технологий.

К настоящему моменту известен широкий перечень угроз информационной безопасности при использовании методов машинного обучения [1], а именно:

1. Угрозы нарушения конфиденциальности данных:
  - 1.1. Извлечение данных о параметрах обученных моделей;
  - 1.2. Извлечение данных об обучающей выборке из обученных моделей;
2. Угрозы нарушения доступности данных:
  - 2.1. Искажение (отравление) обучающей выборки с целью ухудшения качества модели;
3. Угрозы нарушения целостности данных:
  - 3.1. Формирование т. н. состязательных входных данных, некорректно обрабатываемых (например, классифицируемых) моделью.

Модели машинного обучения, предназначенные для решения задач классификации данных, извлекают из входного примера некоторый вектор признаков, который затем используется для определения принадлежности к заданным заранее классам. Способы отнесения вектора признаков к тому или иному классу разнятся в зависимости от архитектуры модели машинного обучения. Например, одним из подходов является выделение на этапе обучения эталонного вектора признаков каждого класса. В процессе эксплуатации системы получаемые вектора признаков сравниваются с эталонными векторами и, соответственно, входные данные классифицируются в зависимости от степени близости вектора признаков к одному эталонных векторов. Однако при такой постановке задачи узким местом является необходимость постоянного хранения эталонного вектора в памяти системы, что может приводить к появлению угрозы их утечки с последующим восстановлением соответствующих им входных данных обучающей выборки.

Известно научное направление, связанное с построением защищенных (робастных) нейросетевых моделей, целью которых является встраивание защиты непосредственно в алгоритм ее работы. Технически это реализуется через выбор определенной архитектуры модели, особого алгоритма обучения и специально подобранных входных данных обучающей выборки. В большинстве случаев, как показывает практика, такой подход позволяет усложнить реализацию атак, но не может обеспечить гарантированную защиту.

Механизм построения защищенного классификатора, описываемый в проекте национального стандарта «Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации», является одним из вариантов построения защищенной нейросетевой модели и предназначен в соответствии с разделами 5 и 6 проекта стандарта [2] для защиты от указанных выше атак.

В рамках настоящей работы показано, что описываемый в проекте стандарта механизм уязвим к типу атак, направленных на извлечение данных об обучающей выборке из обученной модели — атаке проверки принадлежности обучающему множеству [4]. В атаках такого типа нарушитель имеет доступ к обученной модели и набору данных, среди которых, возможно, есть те, которые были использованы для обучения модели. Нарушитель, подавая последовательно на вход модели элементы набора данных, пытается определить на основе значений выходов, принадлежали ли данные обучающему множеству. С теоретической точки зрения возможность построения таких атак связана с наличием у нейросетевых моделей так называемого эффекта запоминания, заключающегося в том, что обученная модель срабатывает на входных данных из обучающей выборки «лучше», чем на произвольных данных. Практическая возможность реализации такой атаки возникает, например, в системах отменяемой биометрии [5] при их реализации на мобильных устройствах. Действительно, необходимость идентификации пользователя по биометрическим признакам без доступа к сети Интернет требует хранения контрольных биометрических шаблонов пользователей на мобильном устройстве, в том числе, в виде защищенных нейросетевых моделей [5], к которым относится и предлагаемый в проекте стандарта механизм. В случае компрометации (утери) устройства пользователем, нарушитель получает возможность непосредственного доступа к хранящейся на устройстве модели и способен проводить атаку проверки принадлежности обучающему множеству перебирая доступные ему, например, из социальных сетей входные биометрические образы потенциальных владельцев устройства.

## **2. Обозначения и краткое описание работы предложенного в проекта стандарта механизма**

Исследуемый способ построения нейросетевых моделей (в терминах стандарта – нейросетевой преобразователь, НКП) предназначен для реализации классификаторов, на вход которых подается предварительно извлеченный из входных данных каким-либо способом вектор признаков, и рассматривает два класса входных признаков:

- «Свой» – класс признаков, которые должны корректно классифицироваться обученным НКП (например, соответствующие изображению лица конкретного человека);
- «Чужие» – произвольные входные признаки, не относящиеся к классу «Свой» (например, соответствующие случайной выборке без возвращения из множества изображений лиц различных людей).

Для целей исследования мы будем также использовать термин «Другой» – выборка из некоторого другого класса, не являющегося «Своим».

- $\overline{a(k)}$  – вектор признаков  $k$ -го входного объекта, если нет двусмысленности, то  $k$  будем опускать;

- $a_j$  –  $j$ -й признак вектора признаков  $\vec{a}$ ;
- $\vec{a}^t$  – вектор мета-признаков;
- $a'_{j^*} - j^*$ - мета-признак вектора мета-признаков  $\vec{a}^t$ ;
- $n$  – количество признаков;
- $m_j$  и  $\sigma_j$  – математическое ожидание и среднеквадратичное отклонение  $j$ -ого признака для класса «Свой»;
- $\mu_j$  и  $\delta_j$  – нормирующие коэффициенты, вычисляемые как математическое ожидание и среднеквадратичное отклонение  $j$ -го признака для класса «Чужие»;
- $K_G$  – количество тренировочных примеров образа «Свой»;
- $K_I$  – количество тренировочных примеров образов «Чужой»;
- $\eta$  – количество входов корреляционного нейрона;
- $\omega_j$  – вес корреляционного нейрона под номером  $j$ ;
- $C_{j,t}$  – коэффициент корреляции между признаками с номерами  $j$  и  $t$ .

Признаки  $\vec{a}(k)$  в соответствии с [2,3] должны иметь распределение, близкое к нормальному, что может быть достигнуто путем подбора и настройки алгоритма извлечения признаков из входных данных, например, вариационного автоэнкодера.

Мета-признаки вычисляются следующим образом:

$$a'_{j^*} = a'_{t,j} = f(a_t, a_j) = \left| \left| \frac{a_t}{\delta_t} \right|^{0.9} - \left| \frac{a_j}{\delta_j} \right|^{0.9} \right|,$$

где

$$\delta_i = \delta_{I,i} = \sqrt{\frac{1}{n} \sum_{K_I} (a_{I,i} - \mu_{I,i})^2},$$

$$\mu_{I,i} = \frac{1}{n} \sum_{K_I} a_{I,i}.$$

Значение нейрона вычисляется следующим образом:

$$y = \sqrt{\frac{1}{\eta} \sum \omega_i (a'_i - m)^2},$$

$$m' = \frac{1}{\eta} \sum_{i=1}^{\eta} a'_i.$$

Выбор мета-признаков при обучении каждого нейрона осуществляется псевдослучайным образом с учетом значения коэффициента корреляции между ними. Конкретные границы на такие значения приведены в тексте проекта стандарта [2], однако они не принципиальны для реализации предлагаемого метода и поэтому опущены в настоящей статье.

Весы синапсов вычисляются следующим образом:

$$\omega_i = \frac{|m''_{G,i} - m''_{I,i}|}{\sigma_{G,i} \sigma_{I,i}},$$

где  $m''_{G,i}, m''_{I,i}$  – математические ожидания, а  $\sigma''_{G,i}, \sigma''_{I,i}$  – среднеквадратичные отклонения значений  $i$ -го мета-признака второго порядка:

$$a''_i = (a'_i - m'')^2.$$

Упрощенно говоря, вес нейрона хранит информацию о расстоянии между центром класса «Свой» и центром класса «Чужие» для соответствующих признаков. Значение функции выхода нейрона имеет вид:

$$\phi(y) = \begin{cases} 3, y < T_{left}; \\ 2, T_{left} \leq y < T_{middle}; \\ 1, T_{middle} \leq y < T_{right}; \\ 0, y \geq T_{right}. \end{cases}$$

Границы  $T_{left}, T_{middle}, T_{right}$  подбираются таким образом, чтобы на выходе выполнялись соотношения  $P(0) \approx P(1) \approx P(2) \approx P(3)$  в предположении о нормальности распределения выхода каждого нейрона. О том, какое именно состояние активации (далее –  $\phi(y)$ ) соответствует гипотезе «Свой», известно только на этапе обучения корреляционных нейронов, злоумышленник не обладает этой информацией, так как она не сохраняется после настройки нейрона.

Далее выбирается хэш-таблица для кодирования полученных значений, но ее выбор также не принципиален для дальнейшего изложения, поскольку она является частью обученного НКП и не является секретной.

Таким образом, обученный преобразователь хранит значения следующих параметров:

- связи корреляционных нейронов с мета-признаками;
- нормирующие коэффициенты признаков  $\delta_i$ ;
- веса нейронов;
- границы  $T_{left}, T_{middle}, T_{right}$ ;
- хэш-таблица для кодирования выходных значений нейронов.

На рис. 1 приведена схема автоматического обучения нейросетевых моделей, представленная в проекте стандарта [3].

### 3. Идея атаки

В основе предлагаемой атаки проверки принадлежности обучающему множеству лежит следующая гипотеза: обучение различных НКП на близких входных данных, например, на различных изображениях одного и того же человека, должно давать в некотором смысле «близкие» обученные нейросетевые преобразователи.

Будем предполагать, что объекты истинного класса «Свой» находятся в выборке, доступной нарушителю. Тогда нарушитель может реализовать атаку с помощью следующей последовательности действий:

- извлечь из атакуемого НКП связи корреляционных нейронов с мета-признаками, границы  $T_{left}, T_{middle}, T_{right}$ , нормирующие коэффициенты признаков  $\delta_i$ ;
- используя каждый доступный ему набор данных (из некоторого класса) в качестве класса «Свой», попытаться обучить новый НКП с фиксированными связями корреляционных нейронов с мета-признаками и нормирующими коэффициентами  $\delta_i$ , полученными из атакуемого НКП;
- выбрать среди полученных границ  $T_{left}, T_{middle}, T_{right}$  те, которые ближе всего в некоторой метрике к соответствующим границам атакуемого (исходного) НКП.

Можно ожидать, что класс входных данных, который дает наиболее близкие к исходным границы, и был использован для обучения атакуемого нейрона изначально. То есть объекты соответствующего класса относятся к истинному классу «Свой» атакуемого нейрона.

Для проверки этой гипотезы рассмотрим далее ситуацию, когда у нарушителя есть множество данных из двух классов: «Свой» и «Другой», а также НКП, обученный изначально на классе «Свой».

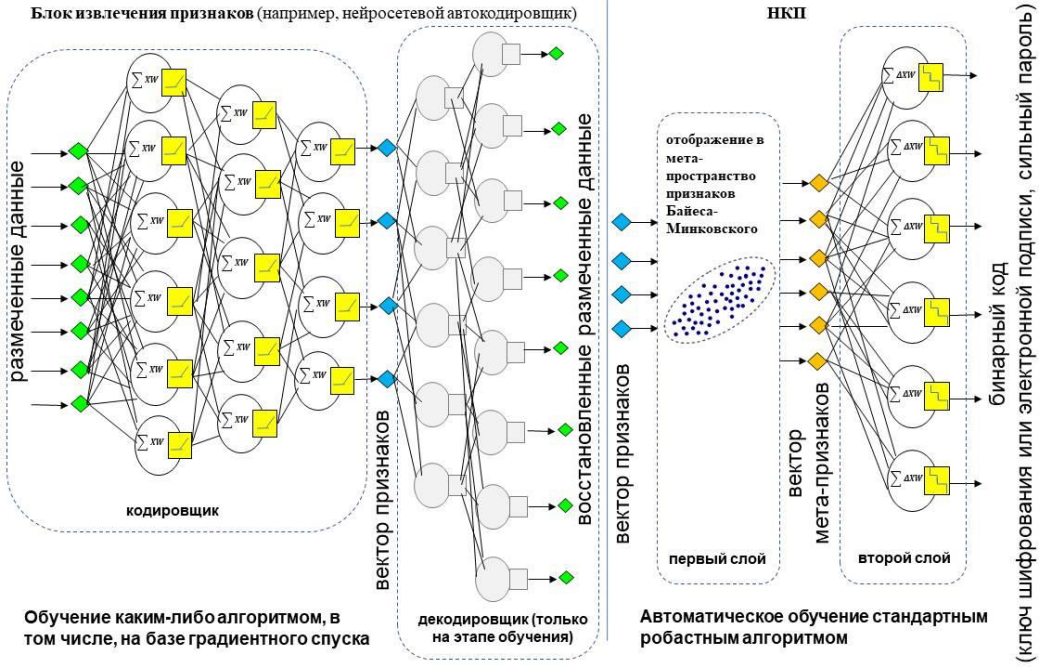


Рис. 1: Схема автоматического обучения нейросетевых моделей  
 Fig. 1: Scheme of automatic training of neural network models

#### 4. Алгоритм атаки

Пусть у нас есть обученный НКП, а значит нам известны связи, нормирующие коэффициенты и веса нейронов. Будем считать, что он был обучен на выборке  $G_1$  примеров «Свой» и выборке  $I_1$  примеров «Чужой».

Сформируем следующие обучающие множества:

- $G_1$  – выборка примеров «Свой»;
- $I_1$  – выборка примеров «Чужой»;
- $G'_1$  – еще одна выборка примеров «Свой» из той же генеральной совокупности, из которой выбиралось множество  $G_1$ ;
- $G_2$  – выборка примеров «Другой» из другой генеральной совокупности;
- $I_2$  – выборка примеров «Чужой» такая, что  $\delta_{I_1,i} = \delta_{I_2,i}$ .

Последнее равенство необходимо для того, чтобы обеспечить близость выборки  $I_1$  к  $I_2$ . В табл. 1 приведены параметры выборок, использовавшихся при практической реализации атаки. Использование нормального распределения определено требованием проекта стандарта [3] о виде распределения входных признаков. Интервал использовавшихся значений математического ожидания выбран в соответствии с параметрами, использованными в контрольном примере [3].

Проект национального стандарта накладывает достаточно жесткие ограничения на распределение входных данных, а именно предполагается, что каждый входной признак имеет нормальное распределение [2] и [3]. Такой эффект действительно наблюдается для определенных вариантов построения НКП, например, тогда, когда на ход к нему поступают данные с вариационного автоэнкодера.

Для проведения экспериментов входные данные формируются в интервале  $[0; 14]$  как указано в контрольном примере процедуры обучения НКП стандарта [2] Для каждой реализации в эксперименте при выработке множеств  $G_1$  и  $G'_1$  для каждого признака фиксируются параметры нормального распределения  $N(\mu_1(i), \sigma_1(i))$ ,  $i = 1, \dots, 128$ , где значение математического ожидания принадлежит указанному диапазону, далее вырабатывается нужное количество векторов признаков. Для множества  $G_2$  фиксируются параметры нормального распределения  $N(\mu_2(i), \sigma_2(i))$ ,  $i = 1, \dots, 128$  с теми же ограничениями, далее вырабатывается нужное количество векторов признаков. Таким образом моделируются выборки векторов из классов «Свой» и «Другой». Как было отмечено ранее выбор вида функции распределения признаков определяется требованиями стандарта [2], однако нетрудно видеть, что предложенный в данной статье метод будет работать и для других классов распределений признаков в случае, если они имеют конечные математические ожидания.

Для множеств  $I_1$  и  $I_2$  при выработке каждого вектора фиксируется  $N(\mu_3(i), \sigma_3(i))$ ,  $i = 1, \dots, 128$ , а затем из него делается выборка одного значения, что моделирует выборку случайных векторов (класс «Чужой»).

Табл. 1. Значение параметров выборки  
Table 1. The values of the parameters of the training sets

Номер эксперимента	Параметры нормального распределения для выборок $G_1$ и $G'_1$	Параметры нормального распределения для выборок $I_1$ и $I'_1$	Параметры нормального распределения для выборок $G_2$
1	$\mathcal{N}(\mu, 1), \mu = 2, 3$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, 3$
2	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 4$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 4$
3	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 5$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 5$
4	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 6$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 6$
5	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 7$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 7$
6	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 8$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 8$
7	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 9$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 9$
8	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 10$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 10$
9	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 11$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 11$
10	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 12$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 13$	$\mathcal{N}(\mu, 1), \mu = 2, \dots, 12$

Тогда эксперимент можно описать следующим образом:

- 1) Сформировать множества  $G_1, G'_1, G_2, I_1$  и  $I_2$ ;
- 2) Обучить НКП-1 на исходных множествах  $G_1$  и  $I_1$ , в результате чего получаются наборы  $T_{left}^1, T_{middle}^1, T_{right}^1$ ;
- 3) Обучить НКП-2 на множествах  $G'_1$  и  $I_2$ , используя связи нейронов с мета-признаками и нормирующие коэффициенты из НКП-1, в результате чего получаются наборы  $T_{left}^2, T_{middle}^2, T_{right}^2$ ;

- 4) Обучить НКП-3 на множествах  $G_2$  и  $I_2$ , используя связи нейронов с мета-признаками и нормирующие коэффициенты из НКП-1, в результате чего получаются наборы  $T_{left}^3, T_{middle}^3, T_{right}^3$ ;
- 5) Вычислить значение статистики  $F$  (описана далее) и сравнить его с граничным значением.

Эксперимент повторяется несколько раз (в представленных далее результатах эксперимент повторялся 50 раз). Таким образом, мы строим статистический критерий различения классов «Свой» и «Другой» относительно обученного НКП.

Обучение НКП-2 и НКП-3 (см. табл. 2) отличается от обучения НКП-1 тем, что мы не оцениваем корреляцию признаков для выбора связей, а берем значения связей из НКП-1. Возможны следующие варианты для каждого нейрона в НКП с номером  $i$ .

Табл. 2. Возможные варианты обучения нейронов в эксперименте  
 Table 2. Possible variants neurons training

Нейрон НКП-2	Нейрон НКП-3
обучен	обучен
обучен	не обучен
не обучен	обучен
не обучен	не обучен

Нас интересуют нейроны, соответствующие первой строке, т.е. которые обучились для всех трех НКП.

Для каждого такого нейрона (пусть их будет  $d$  штук) на шаге 4 вычисляются следующие статистики:

$$z_2(i) = \max(|T_{left}^1(i) - T_{left}^2(i)|, |T_{middle}^1(i) - T_{middle}^2(i)|, |T_{right}^1(i) - T_{right}^2(i)|), i = 1, \dots, d,$$

$$z_3(i) = \max(|T_{left}^1(i) - T_{left}^3(i)|, |T_{middle}^1(i) - T_{middle}^3(i)|, |T_{right}^1(i) - T_{right}^3(i)|), i = 1, \dots, d.$$

Далее вычисляется результирующая статистика:

$$F = \sum_{i=1}^d I\{z_2(i) < z_3(i)\}$$

Если полученное значение статистики стабильно больше  $\frac{d}{2}$  для всех проведенных экспериментов, то гипотеза верна, и мы можем реализовать атаку проверки принадлежности обучающему множеству. В табл. 3 приведены средние значения статистики  $F > \frac{d}{2}$  для каждого эксперимента, которые показывают успешность предложенной атаки.

## 5. Заключение

Для алгоритма обучения нейросетевого преобразователя из проекта национального стандарта «Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации» предложен статистический критерий, который позволяет реализовать атаку проверки принадлежности обучающему множеству.

Проведены экспериментальные исследования, которые показали, что на входных данных, соответствующих ограничениям, приведенным в проекте стандарта, критерий позволяет корректно определять входные данные, использовавшиеся для обучения атакуемого нейросетевого преобразователя.



Табл. 3. Результаты эксперимента

Table 3. Experimental results

Номер эксперимента	Количество «успешно» обученных нейронов НКП-2	Количество «успешно» обученных нейронов НКП-3	Среднее значение статистики $F$	Число запусков, в которых значение статистики $F > \frac{d}{2}$
1	128	128	68	45
2	128	128	78	48
3	128	128	81	49
4	128	128	72	46
5	128	128	91	50
6	128	128	88	50
7	128	128	83	49
8	128	128	77	48
9	128	128	65	46
10	128	128	93	50

## Список литературы / References

- [1]. Гуселев А.М., Маршалко Г.Б., «Проблемы безопасности систем машинного обучения», в сборнике трудов МИТСОБИ'2021, стр. 23-24.
- [2]. Первая редакция проекта национального стандарта «Искусственный интеллект. Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации», Технический комитет по стандартизации «Искусственный интеллект» (ТК 164), 2022.
- [3]. Сулавко, А.Е. «Защищенный режим исполнения искусственного интеллекта на базе автоматически обучаемых сетей автокорреляционных нейронов», Технический отчет, ОмГТУ, Омск, 2021, 101 с.
- [4]. R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership Inference Attacks against Machine Learning Models, 2017 IEEE Symposium on security and privacy (SP), pp. 3-18, 2017.
- [5]. Manisha, N. Kumar, Cancelable biometrics: a comprehensive survey, Artificial intelligence review, 53, 3403-3446, 2020.

## Информация об авторах / Information about authors

Григорий Борисович МАРШАЛКО – эксперт, Технический комитет по стандартизации "Криптографическая защита информации". Область научных интересов: защита информации, криптография, биометрическая идентификация.

Grigory Borisovich MARSHALCO – Expert, Technical committee for standardization "Cryptography and security mechanisms". Research interests: information security, cryptography, biometric identification.

Роман Александрович РОМАНЕНКОВ – эксперт, Технический комитет по стандартизации "Криптографическая защита информации". Область научных интересов: защита информации, моделирование случайных величин, прикладные методы математической статистики.

Roman Alexandrovich ROMANENKOV – Expert, Technical committee for standardization "Cryptography and security mechanisms". Research interests: information security, cryptography, modeling of random variables, applied mathematical statistics.

Юлия Анатольевна ТРУФАНОВА – эксперт, Технический комитет по стандартизации "Криптографическая защита информации". Область научных интересов: защита информации, биометрическая идентификация, машинное обучение.

Julia Anatolievna TRUFANOVA – Expert, Technical committee for standardization "Cryptography and security mechanisms". Research interests: information security, biometric identification, machine learning.