конференция РусКрипто

Защищенное исполнение нейросетевых алгоритмов классификации образов для задач биометрической аутентификации на базе сетей корреляционных нейронов

Работа выполнена ОмГТУ в рамках государственного задания Минобрнауки России на 2023-2025 годы № FSGF-2023-0004

Сулавко Алексей Евгеньевич
д.т.н., доцент кафедры комплексной защиты информации, главный научный сотрудник, **ФГАОУ ВО ОмГТУ**

Актуальность исследования

Объем мирового рынка биометрических систем

Международное аналитическое агентство **Precedence** Research, 2030 год

Прогноз объемов мирового рынка от **Transparency Market Research** на 2031 год:

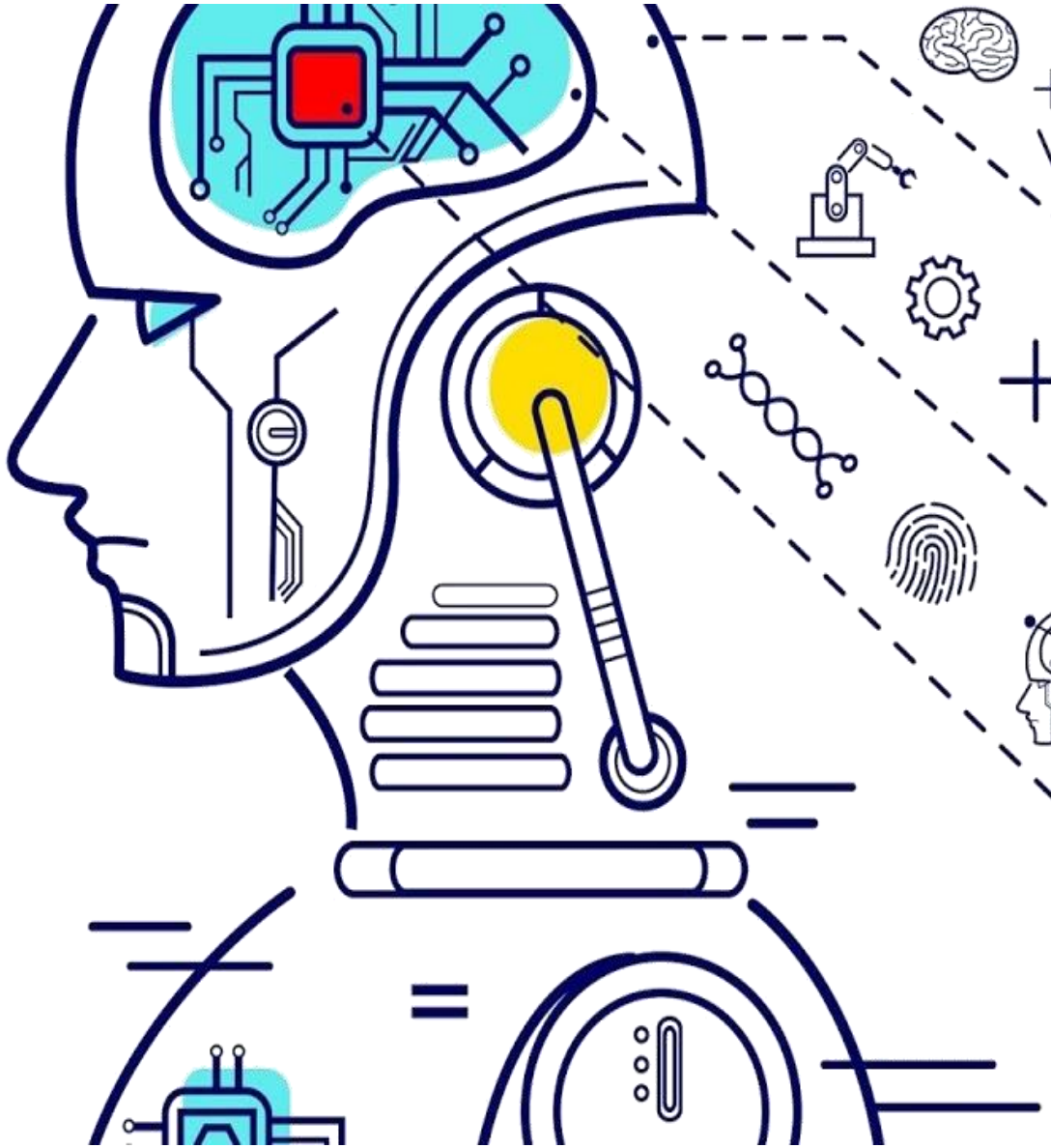
- рынок биометрии достигнет объема $136,18 млрд. (**CAGR** = 13,3%)
- сегмент биометрических платежных систем достигнет $8,8 млрд. (**CAGR** = 56,2%)
- сегмент лицевой биометрии достигнет $20,5 млрд.
- сегмент бесконтактной и многофакторной биометрической аутентификации – $59,5 млрд.
- крупные инвестиции направлены на сегменты **голосовой** и **лицевой** биометрии

Ключевые проблемы

- Подверженность компьютерным атакам
- Компрометация знаний ИИ (биометрических образов)
- Сложность автоматизации обучения на малых выборках
- Низкая робастность обучения
- Недостаточная надежность*

*недостаточные показатели вероятностей «ложного отказа» **FRR** и «ложного допуска» **FAR** или неспособность системы сохранять их с течением времени на приемлемом уровне

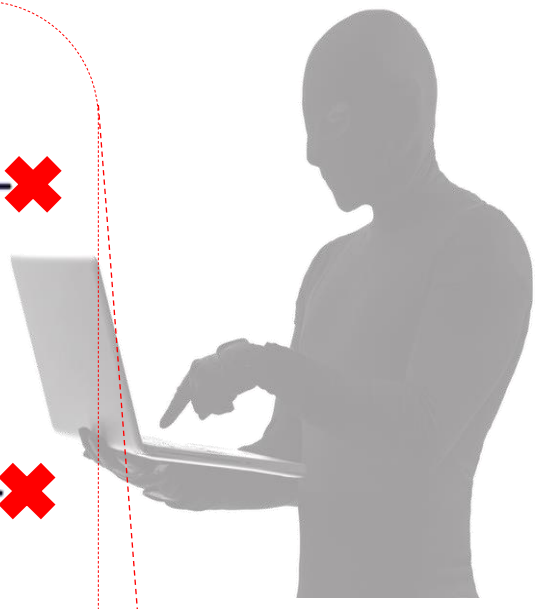
Защищенное исполнение нейросетевых алгоритмов доверенного ИИ



Анализ логики работы и операций, совершаемых ИИ

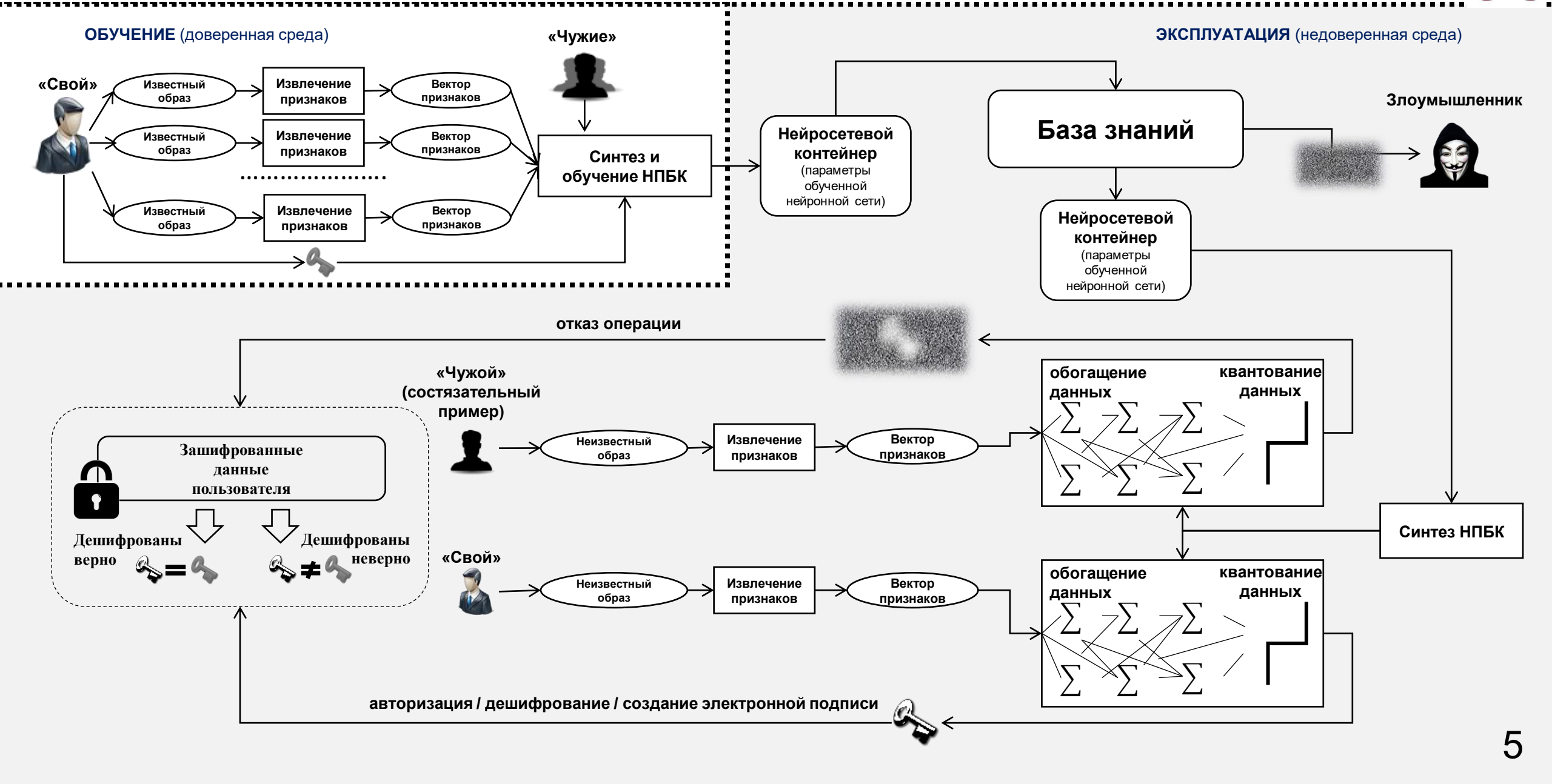
Несанкционированное управление, манипуляции с моделями ИИ

Извлечение и интерпретация знаний из памяти ИИ



1. **НПБК**
2. **Мета-признаки Байеса-Минковского**

Нейросетевые преобразователи биометрия-код (НПБК)



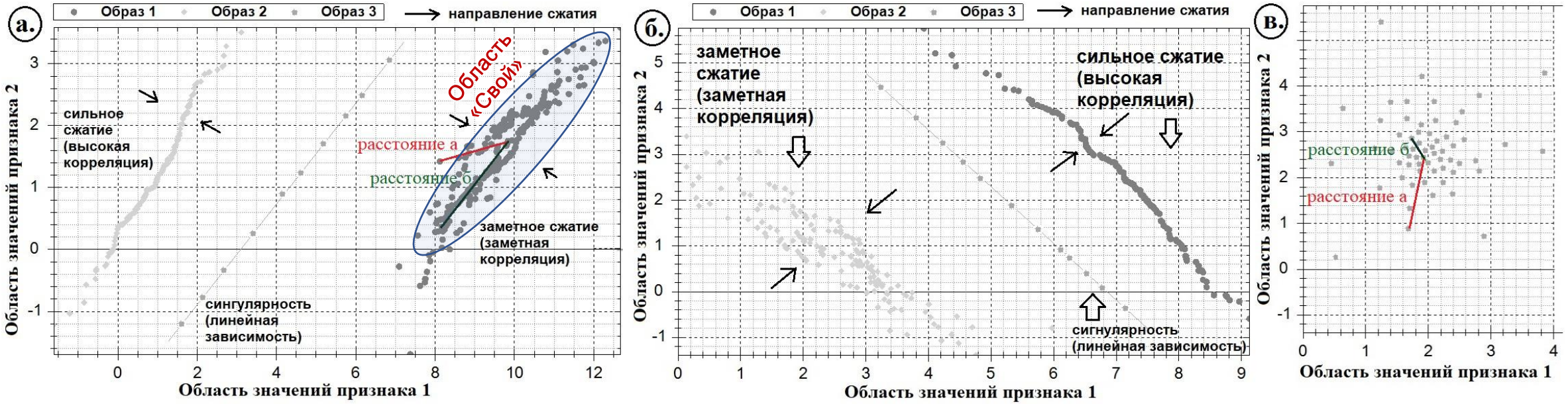
Искривление пространства признаков из-за корреляции между измерениями

конференция РусКрипто

а. при положительной корреляции между признаками

б. при отрицательной корреляции

в. при независимости признаков



$$y = \sqrt[g]{\sum_{j=1}^{n} \left| \frac{(m_j - a_j)}{\sigma_j} \right|^g},$$

где a_j – значение j-го признака; n – количество признаков;

m_j и σ_j – математическое ожидание и среднеквадратичное отклонение значений j-го признака для класса **«Свой»**, с которым сравнивается образ $\bar{a}$ (класс «Свой» представляет биометрические образы одного из легитимных пользователей);

g – степенной коэффициент, определяющий уровень «искривления» пространства.

Метрики Байеса-Минковского

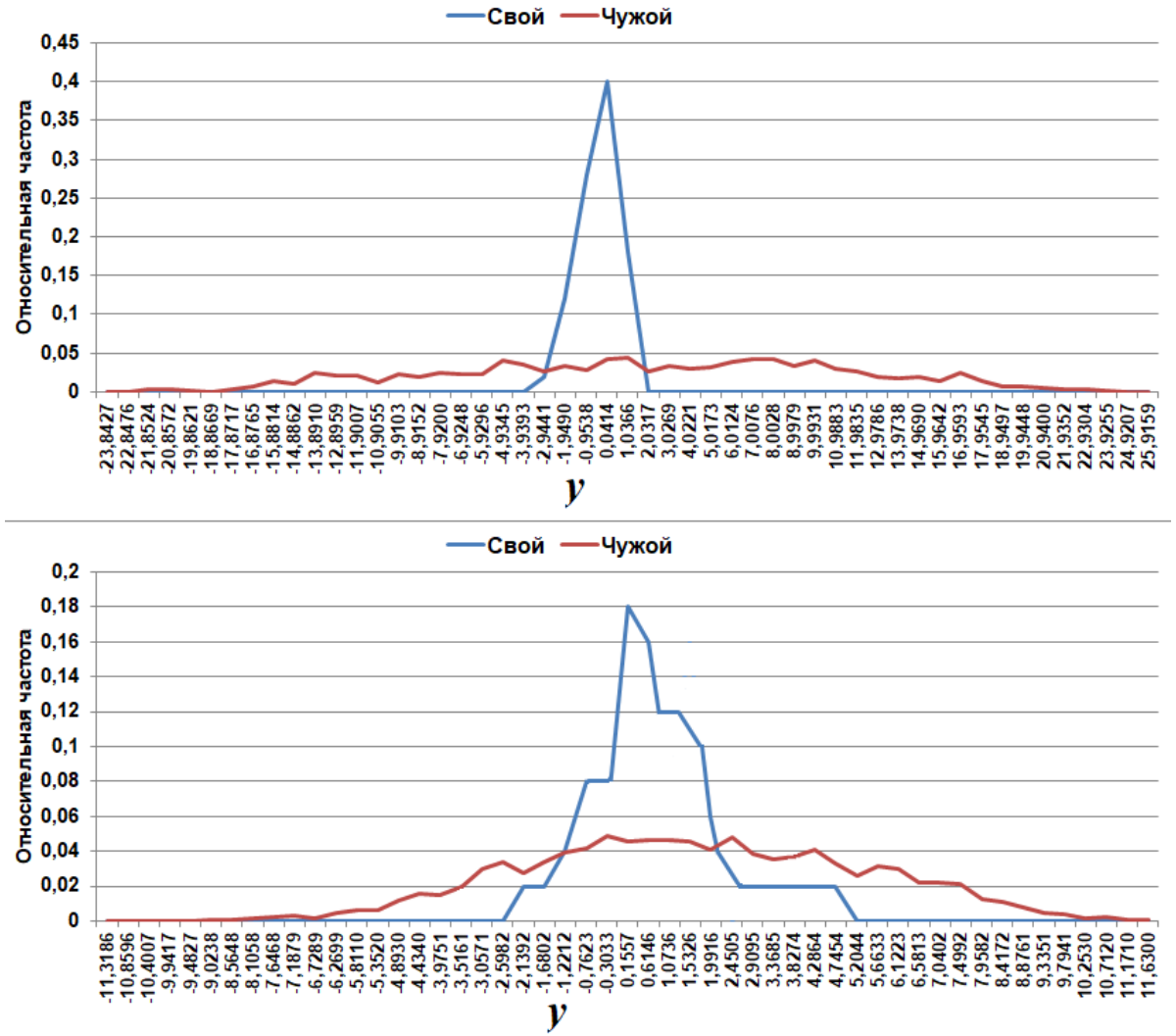


Рисунок – Графики относительных частот значений **меры (1)** при $g = 1$, $n' = 5$:

вверху – для случая сильно зависимых признаков ($C_{j,t} > 0{,}7$);

внизу – для случая независимых признаков ($|C_{j,t}| < 0{,}3$).

$$y = \sqrt[g]{\sum_{j=1}^{n} \left| \left| \frac{m_t - a_t}{\sigma_t} \right|^g - \left| \frac{m_j - a_j}{\sigma_j} \right|^g \right|}, j \neq t, \quad (1)$$

$$y = \sum_{j=1}^{n} \left| \left| \frac{a_t}{\delta_t} \right|^g - \left| \frac{a_j}{\delta_j} \right|^g \right|, j \neq t, \quad (2)$$

$$y = \sqrt{\frac{1}{n} \sum_{j^*=1}^{n} \left(a'_{j^*} - m'\right)^2}, m' = \frac{1}{n} \sum_{t^*=1}^{n} a'_{t^*} \quad (3)$$

где a_j – значение j-го признака;

n – количество признаков;

m_j и σ_j – математическое ожидание и среднеквадратичное отклонение значений j-го признака для класса **«Свой»**, с которым сравнивается образ $\bar{a}$;

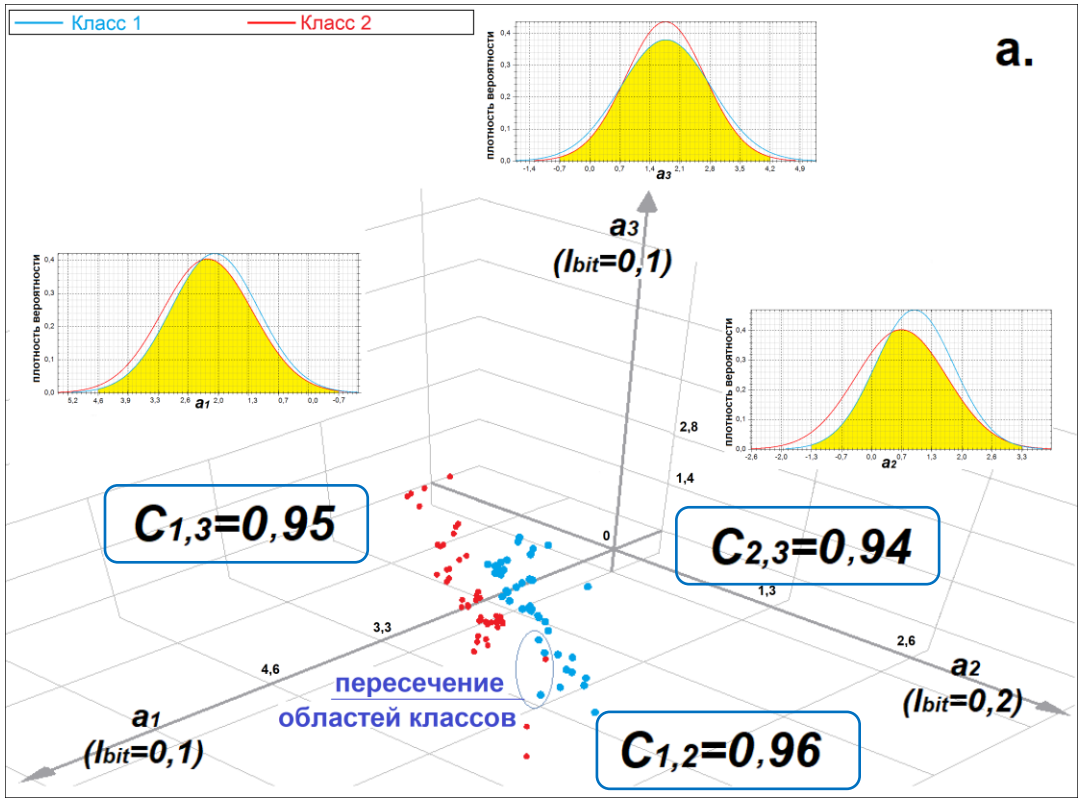
μ_j и δ_j – это нормирующие коэффициенты, вычисляемые как мат. ожидание и стандартное отклонение значений признака для класса **«Чужие»**;

g – степенной коэффициент, определяющий уровень «искривления» пространства.

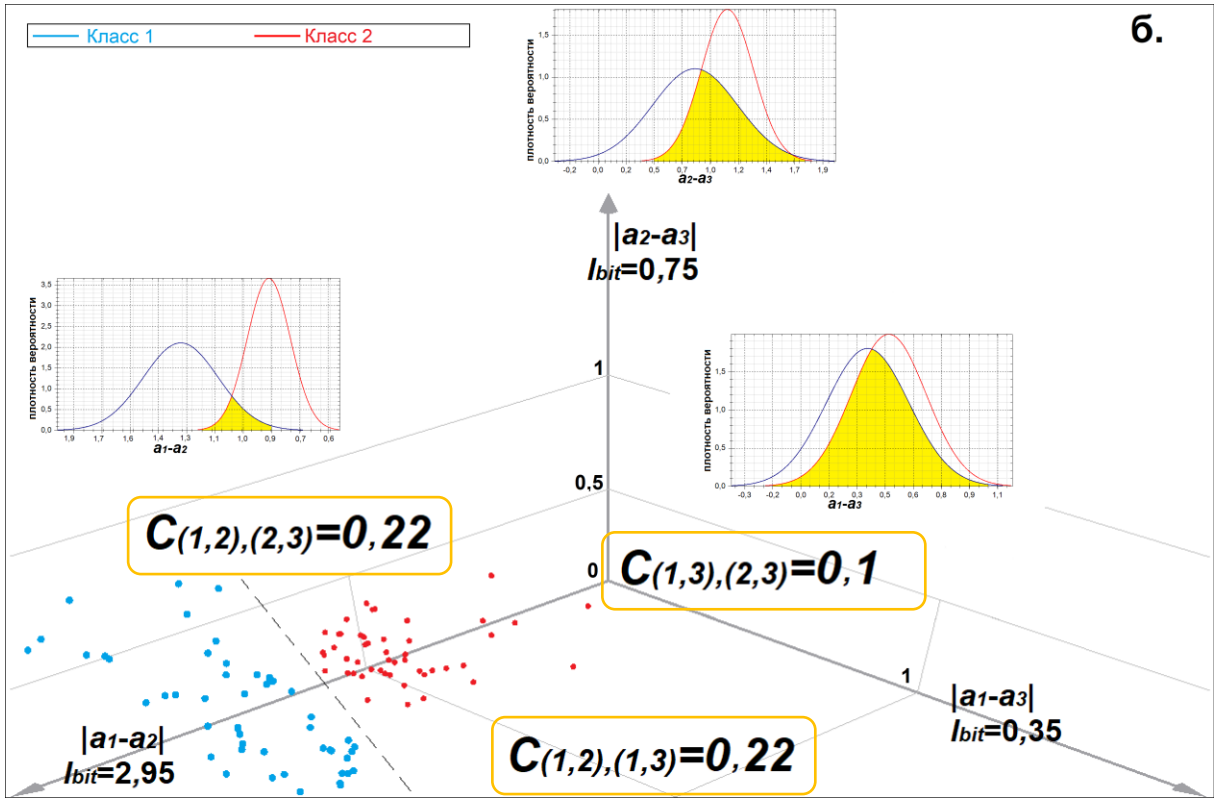
Пространства признаков и мета-признаков Байеса-Минковского

конференция РусКрипто

а. положительно коррелированных признаков



б. мета-признаков, производных от исходных признаков



$$a'_{j*} = a'_{t,j} = f(a_t, a_j) = \left| \left| \frac{a_t}{\delta_t} \right|^g - \left| \frac{a_j}{\delta_j} \right|^g \right|, \quad (1)$$

$$j > t,\ j* = \sum_{t*=1}^{t-1} (n - t*) + j - t \quad (2)$$

$$n' = 0{,}5(n(n\text{-}1)) = 0{,}5n^2 - 0{,}5n \quad (3)$$

где δ_j – это нормирующий коэффициент, вычисляемый как среднеквадратичное отклонение значений j-го признака для класса «Чужие»

$C_{j,t}$ – это коэффициент парной корреляции между признаками под номерами j и t

Модель корреляционного нейрона первого поколения

$$y = \sqrt{\frac{1}{\eta}\sum_{\iota=1}^{\eta} w_\iota \left(a'_\iota - m'\right)^2}, m' = \frac{\sum_{\iota^*=1}^{\eta} a'_{\iota^*}}{\eta} \quad (1)$$

η – количество входов нейрона, w_ι – вес ι-го синапса ($w_\iota > 0$)

$$w_\iota = \frac{|m''_{(G),\iota} - m''_{(I),\iota}|}{\sigma''_{(G),\iota} \cdot \sigma''_{(I),\iota}}, \quad (2)$$

$m_{(G),\iota}''$, $m_{(I),\iota}''$ – математические ожидания,

$\sigma_{(G),\iota}''$, $\sigma_{(I),\iota}''$ – стандартные отклонения ι-го мета-признака

2-го порядка **(3)** для образов **«Свой»** и **«Чужие»**, соответственно

$$a''_\iota = (a'_\iota - m')^2 \quad (3)$$

После обучения параметры $m_{(G),\iota}''$, $m_{(I),\iota}''$, $\sigma_{(G),\iota}''$, $\sigma_{(I),\iota}''$ удаляются.

Активация нейрона (4):

$$\phi(y) = \begin{cases} 3, y < T_{left} \\ 2, T_{left} \le y < T_{middle} \\ 1, T_{middle} \le y < T_{right} \\ 0, y \ge T_{rightt} \end{cases}, \quad (4)$$

T_{left}, T_{middle} и T_{right} – левый, средний и правый **пороги** активации

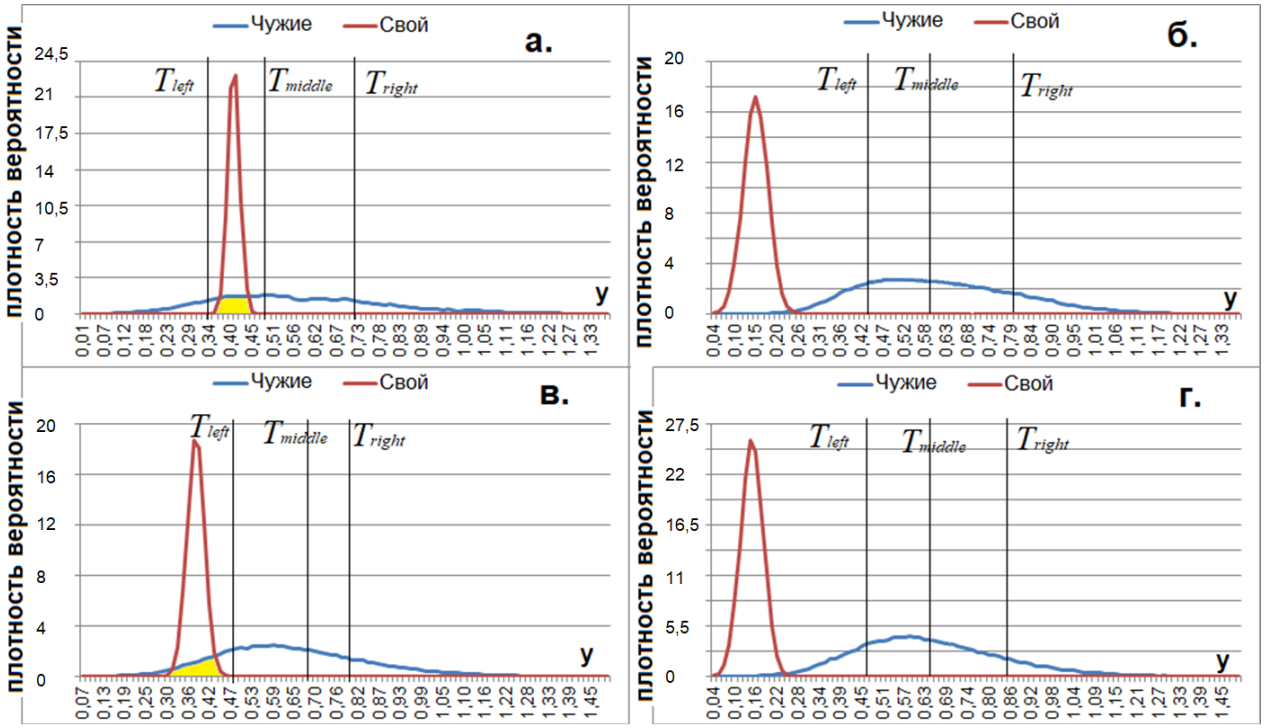


Рисунок – Графики плотностей вероятности значений **меры (1)** для сгенерированных данных при g=1, I≈1,75 бит:

а. для всех классов 1> $C_{j,t}$ >0,95, n'=10;

б. для всех классов -1< $C_{j,t}$ < -0,95, n'=10;

в. для классов «Свой» 1> $C_{j,t}$ >0,95, для класса «Чужие» $|C_{j,t}|$ < 0,3, n'=10;

г. для классов «Свой» -1< $C_{j,t}$ < -0,95, для класса «Чужие» $|C_{j,t}|$ < 0,3, n'=10;

Преимущества и недостатки
модели корреляционных нейронов первого поколения

конференция РусКрипто

Преимущества по сравнению с ГОСТ Р 52633.5	Гипотетический недостаток
Более высокая длина ключа (по результатам экспериментов может многократно превышать длину ключа для НПБК, обученного по ГОСТ Р)	**Заявлена потенциальная атака**, обнародованная в рамках РусКрипто 2023 (авторы: Маршалко Г.Б., Романенков Р.А., Труфанова Ю.А., ТК26)
Более высокая надежность решений (ниже FRR и FAR)	Нейрон не поддерживает работу **со слабо коррелированными парами признаков** (необходимо наличие достаточного количества сильно коррелированных признаков)

На базе первой модели корреляционных нейронов создана первая редакция стандарта ГОСТ Р «Искусственный интеллект. Нейросетевые алгоритмы в защищённом исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации»

Нейро-корреляционный преобразователь (НКП) – нейросетевой преобразователь образов в код на основе сети корреляционных нейронов

Несостоятельность атаки на НКП на базе корреляционных нейронов предыдущего поколения

конференция РусКрипто

1. **Неверное допущение:**
Предполагается, что у нарушителя есть набор примеров из различных классов «Свой», при этом среди них есть один, который использовался для обучения доступного нарушителю НКП

2. **Некорректный критерий сходства между двумя НКП** и, как следствие, некорректные действия атакующего:

- Атакующий пытается обучить несколько НКП путем вычисления весовых коэффициентов и нахождения пороговых значений (границ) T_{left}, T_{middle}, T_{right}, **не учитывая номера хеш-таблиц**.

- Далее он пытается выбрать среди обученных НКП тот, у которого границы ближе всего по некоторой метрике к границам атакуемого нейрона. При этом **различие весовых коэффициентов двух НКП не берется во внимание**.

- **Атакующий строит метрику сходства НКП целиком на близости границ**, а **нужно строить на сходстве выходных состояний нейронов**.

Схожесть границ T_{left}, T_{middle}, T_{right} у двух НКП **не свидетельствует о том**, что у этих НКП область «Свой» будет находиться на идентичных интервалах:

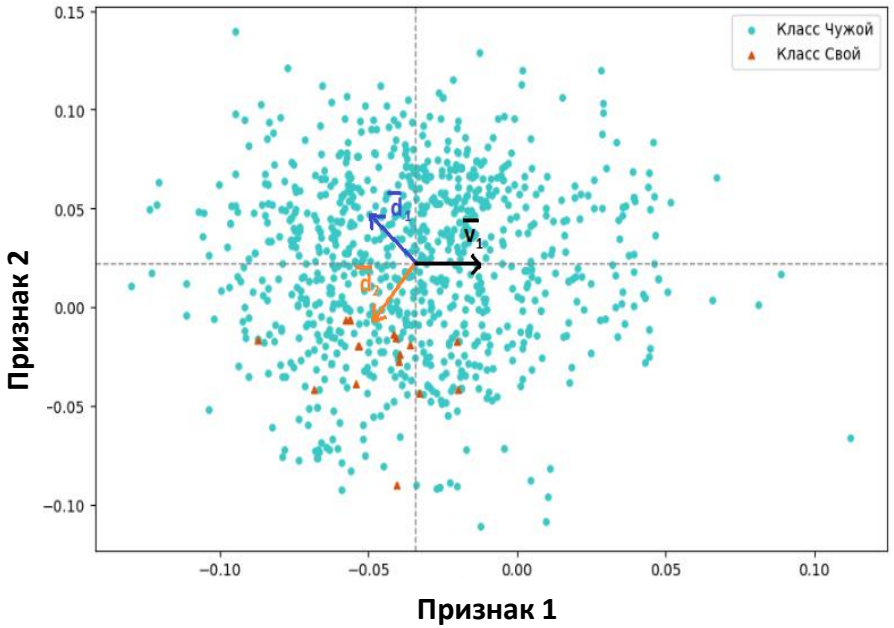
$$[-\infty; T_{left}], \quad [T_{left}; T_{middle}], \quad [T_{middle}; T_{right}], \quad [T_{right}; \infty]$$

В описании атаки не приводится свидетельств того, что фальсифицированный и легитимный НКП будут генерировать схожие или идентичные выходные коды в ответ на одинаковые входные образы.
Не показана корреляция между значением критерия и правильным выбором интервала.

3. **Не понятно, какое количество нейронов использовано в рамках атаки на НКП.**

Итог: выводы авторов преждевременны, предложенный авторами критерий не позволяет корректно определять входные данные, использовавшиеся для обучения атакуемого НКП.

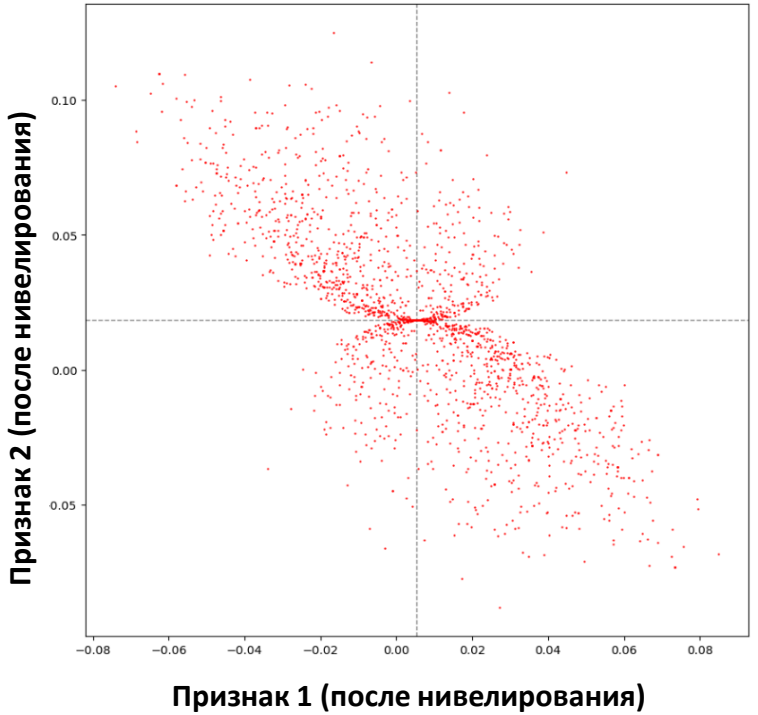
Драфт второй модели корреляционного нейрона для работы со слабо коррелированными признаками (разработана **Панфиловой И.Е.**)



Евклидовы расстояния от центра «Все Чужие» до образа определенного класса (d_2) и случайного образа (d_1) в пространстве исходных признаков

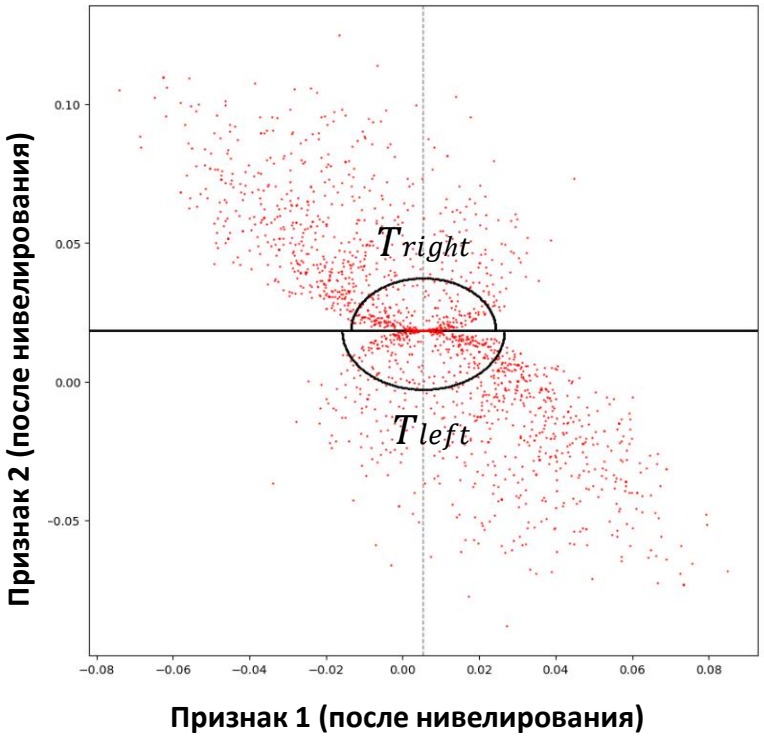
$$a'_{i,j} = \sqrt{(a_i - m_i)^2 + (a_j - m_j)^2} * sin\left(\widehat{\bar{d}, \bar{v}}\right) \quad (1)$$

тригонометрическое нивелирование

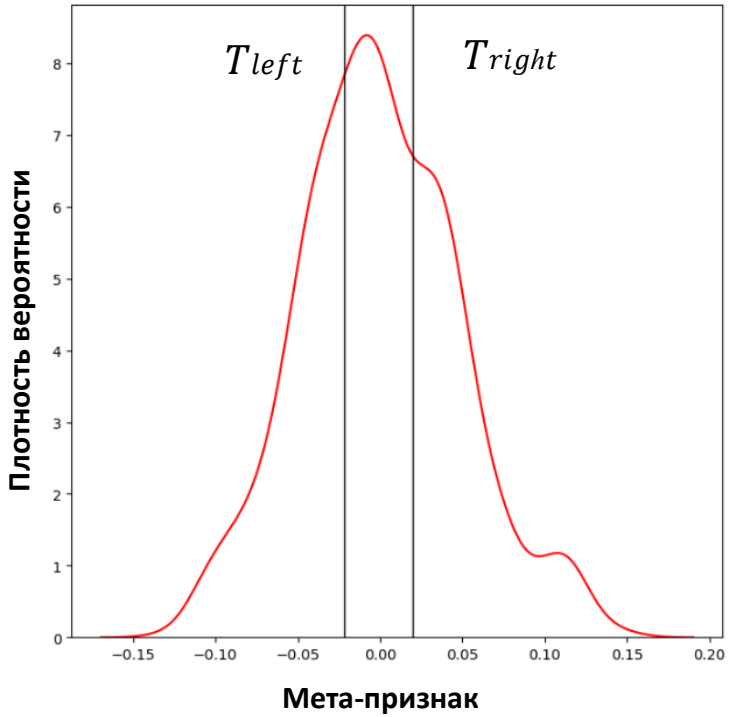


Мета-пространство, порождаемое метрикой (1)

Драфт второй модели корреляционного нейрона для работы со слабо коррелированными признаками (разработана **Панфиловой И.Е.**)



Визуализация порогов
в мета-пространстве признаков



Визуализация порогов на графике функции плотности вероятности распределения мета-признаков (1)

Евклидово расстояние, скорректированное путем тригонометрического нивелирования, – это **новый мета-признак (1)**

$$a'_{i,j} = \sqrt{(a_i - m_i)^2 + (a_j - m_j)^2} * sin\left(\widehat{\vec{d}, \vec{v}}\right) \qquad (1)$$

Драфт второй модели корреляционного нейрона для работы со слабо коррелированными признаками (разработана **Панфиловой И.Е.**)

Функционал нейрона:

$$y = \sum_{i=1}^{n} \sqrt{(a_i - m_i)^2 + (a_j - m_j)^2} * sin(\widehat{\bar{d}, \bar{v}}), i \neq j \qquad (2)$$

Функция активации нейрона:

$$\phi(y) = \begin{cases} 1, y < T_{left} \\ 0, T_{left} \leq y < T_{right}, \\ -1, y \geq T_{rightt} \end{cases} \qquad (3)$$

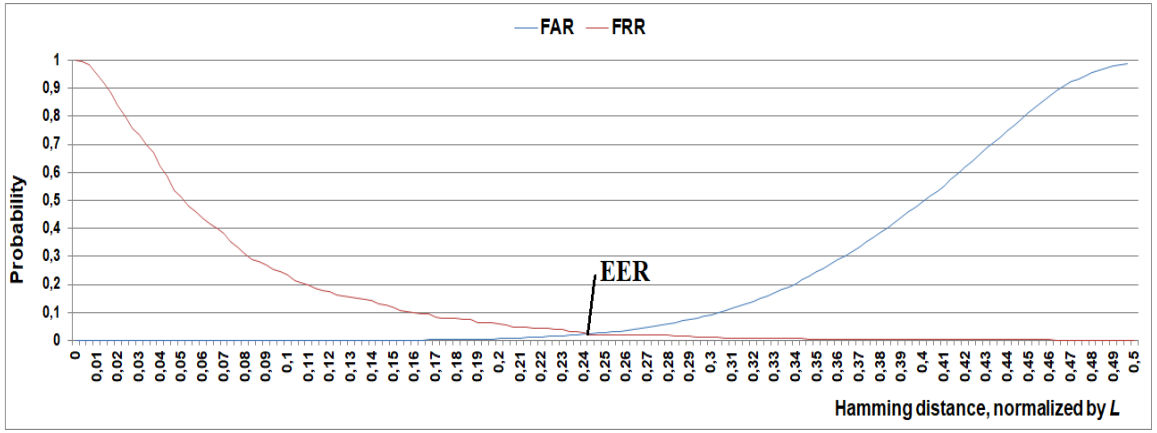
Хеш-таблица №	-1	0	1	№	-1	0	1
1	11	00	01	13	01	00	11
2	11	00	10	14	01	00	10
3	11	01	00	15	01	10	00
4	11	01	10	16	01	10	11
5	11	10	00	17	01	11	10
6	11	10	01	18	01	11	00
7	00	01	11	19	10	00	01
8	00	01	10	20	10	00	11
9	00	10	01	21	10	01	11
10	00	10	11	22	10	01	00
11	00	11	10	23	10	11	00
12	00	11	01	24	10	11	01

Предварительные результаты экспериментов с моделями (выполнено Панфиловой И.Е.)

конференция РусКрипто

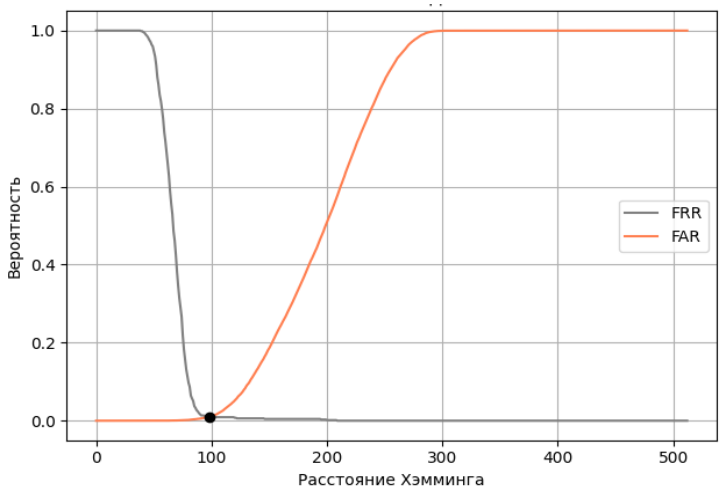
Первая модель
(распознавание по эхограммам ушного канала):

AIC-ears-75 (EER = 0,0238%):

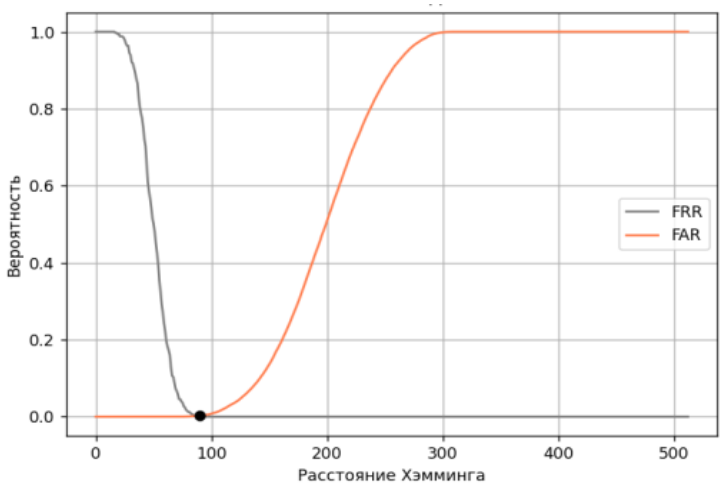


Вторая модель
(распознавание по лицу):

Labeled Faces in the Wild (EER = 0,012%):

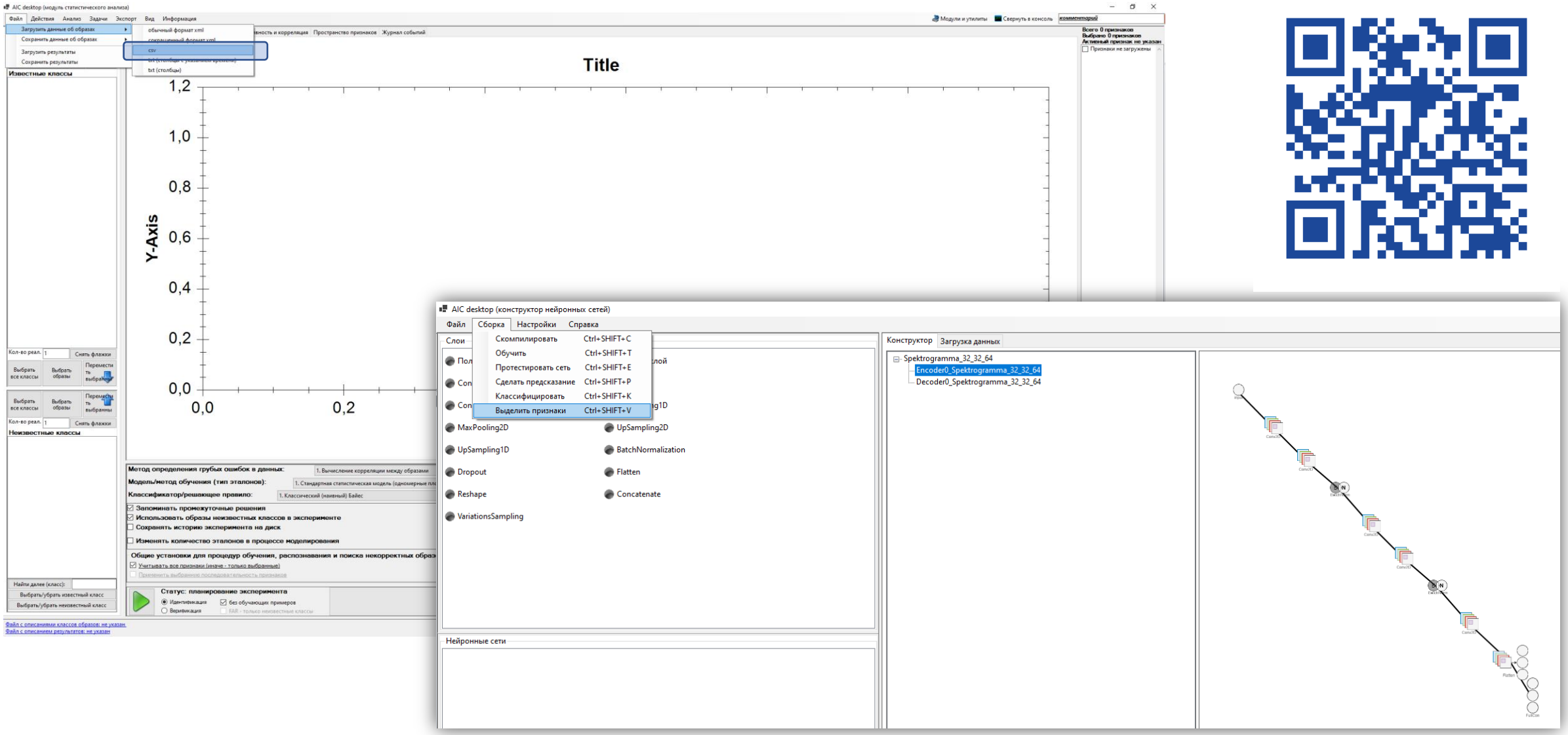


Faces94 (EER = 0,003%):



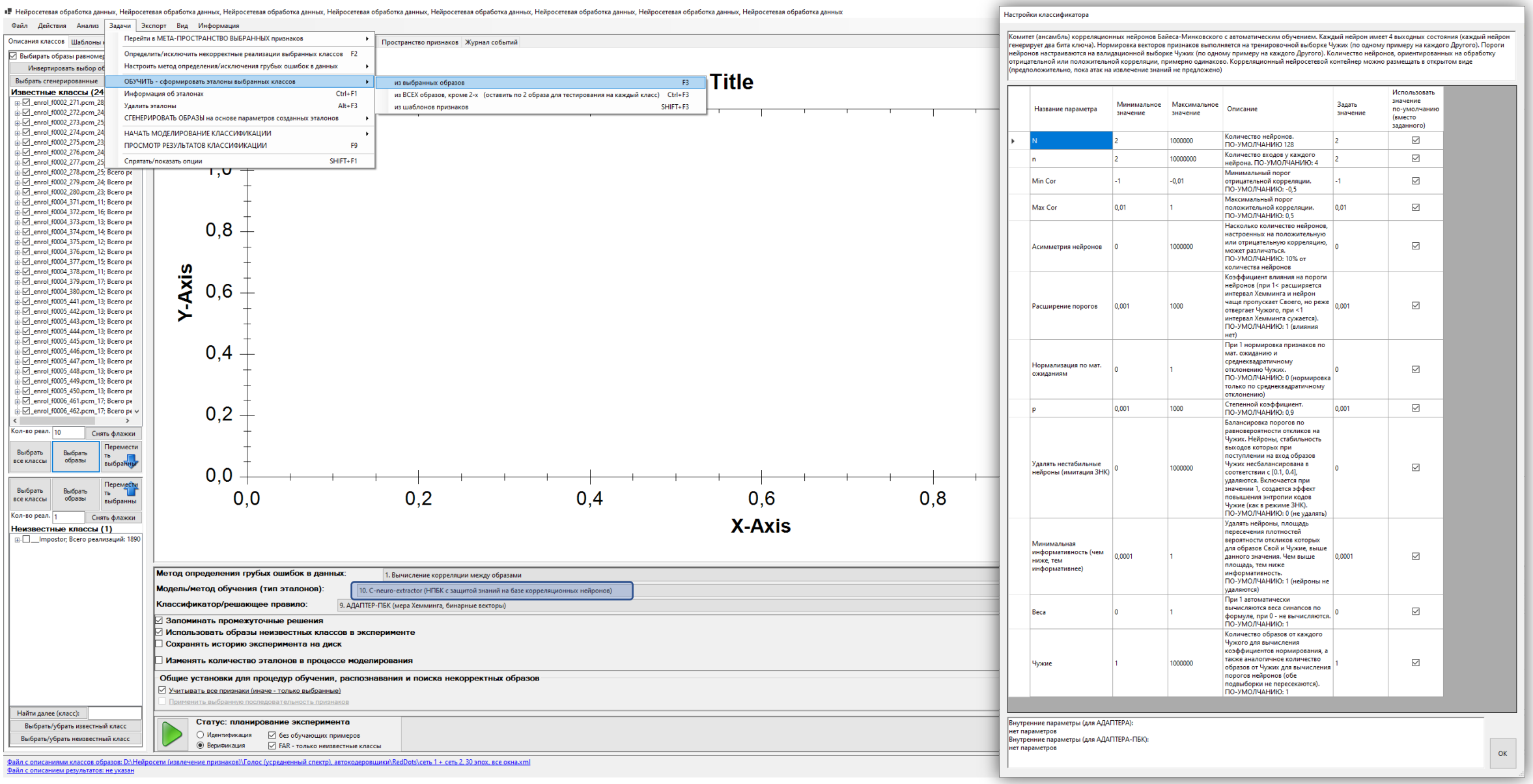
AIC desktop

конференция РусКрипто



AIC desktop

конференция РусКрипто



конференция РусКрипто

Подробнее о направлениях наших исследований вы можете узнать на сайте AIC desktop

AIC desktop

Алексей Сулавко
(ОмГТУ), д.т.н.

AIC Platform

Контакты:

sul@aiconstructor.ru
+7 953 394-90-54

Оценка информативности мета-признаков Байеса-Минковского

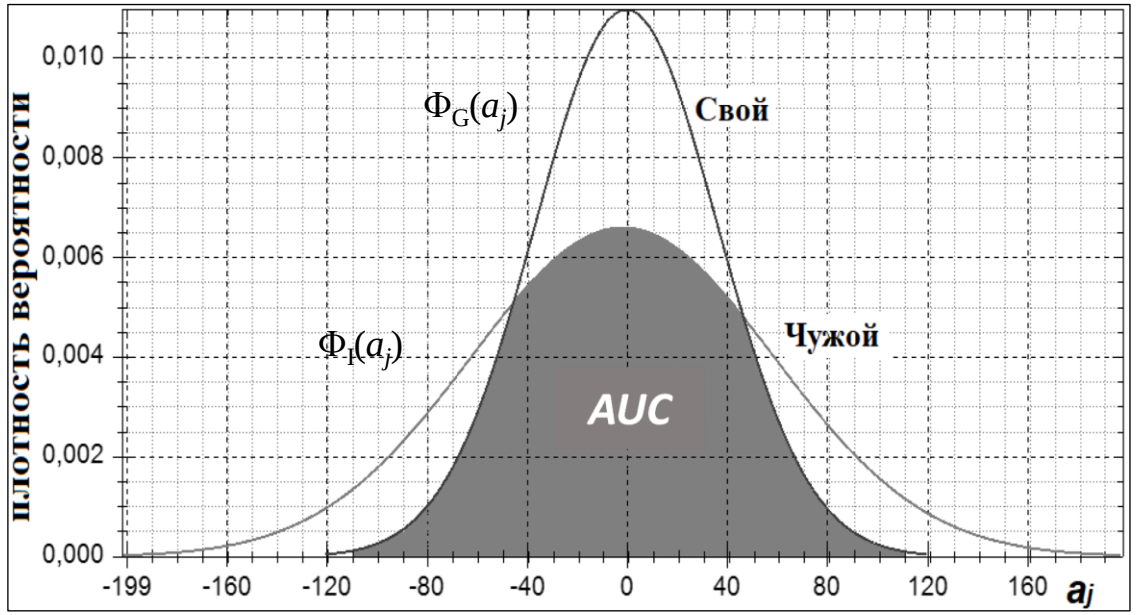


Рисунок 1 – Определение информативности признака/мета-признака для одного из классов образов

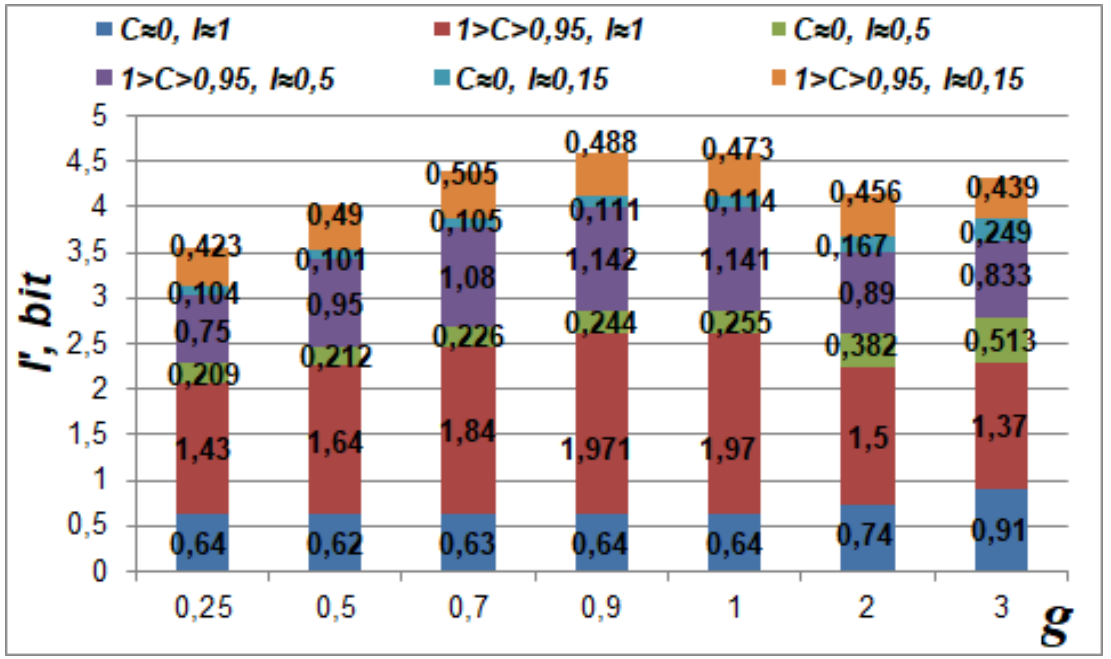


Рисунок 2 – Информативность мета-признаков I'

Мета-признак:

$$a'_{j^*} = a'_{t,j} = f(a_t, a_j) = \left| \left| \frac{a_t}{\delta_t} \right|^g - \left| \frac{a_j}{\delta_j} \right|^g \right|, \quad (1)$$

$$j > t,\ j^* = \sum_{t^*=1}^{t-1} (n - t^*) + j - t \quad (2)$$

где δ_j – это нормирующий коэффициент, стандартное отклонение значений признака для класса «Чужие»

Информативность:

$$I_j = -\log_2(AUC(\Phi_G(a_j), \Phi_I(a_j))), \quad (3)$$

где AUC – площадь, ограниченная функциями плотности вероятности $\Phi_G(a_j)$, $\Phi_I(a_j)$ и осью абсцисс

$\Phi_G(a_j)$ и $\Phi_I(a_j)$ характеризуют значения признака для классов образов («Свой») и «Чужие»

Алгоритм синтеза и обучения корреляционного нейрона первого поколения

При настройке порогов T_{left}, T_{middle} и T_{right} алгоритм обеспечивает:

$0{,}1<P(\phi(y))<0{,}4$, где $P(\phi(y))$ – это относительная частота появления $\phi(y)$ при поступлении на вход образа «Чужой»;

$\phi(y)$ – функция активации нейрона

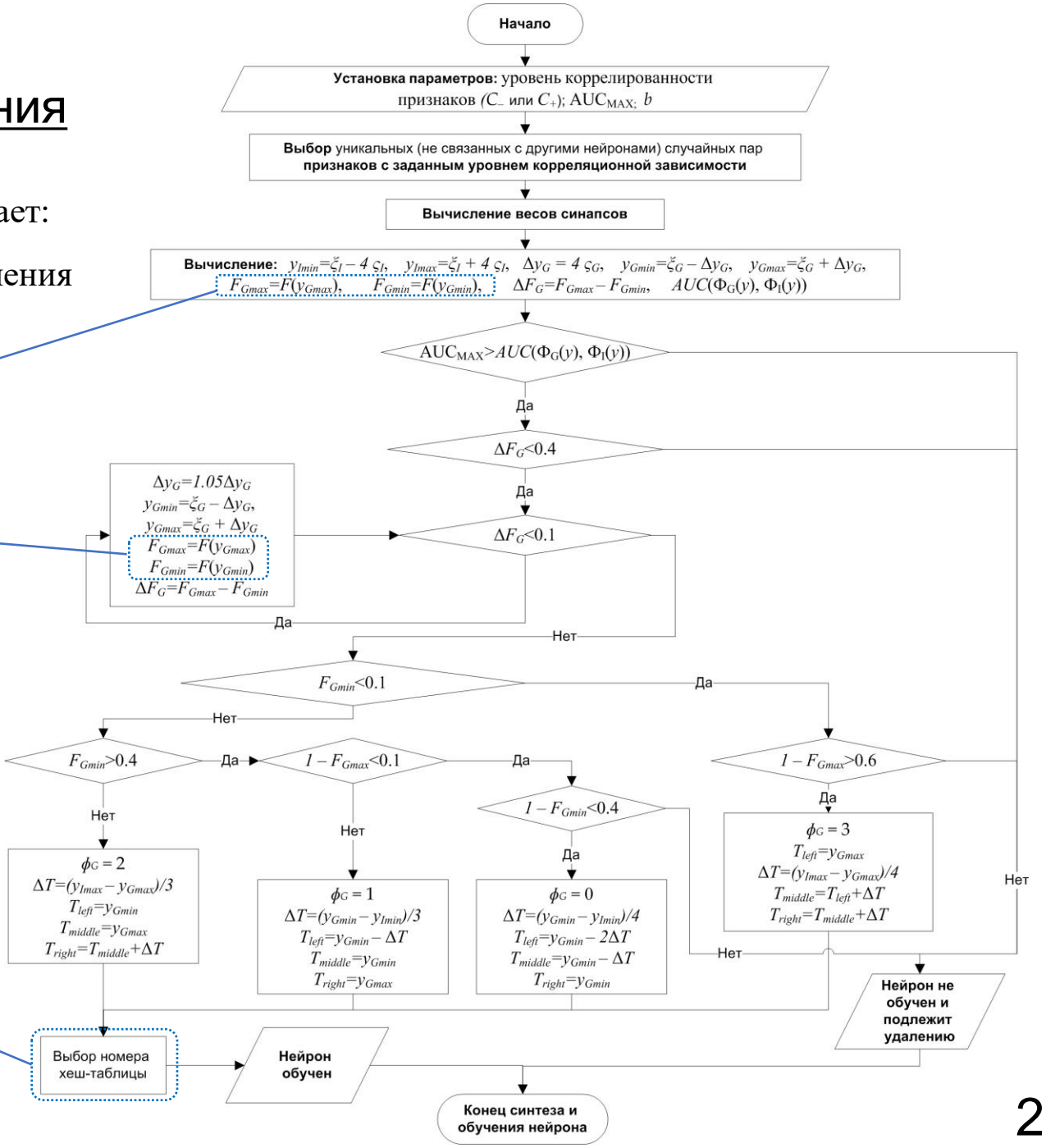
$$F(y)=\int_{-\infty}^{y}\frac{1}{\varsigma\sqrt{2\pi}}e^{-\frac{(q-\xi)^2}{2\varsigma^2}}dq,$$

ξ и ς – математическое ожидание и стандартное отклонение y при поступлении на входы обучающих примеров

Номер преобразования $\phi(y)\rightarrow b$ выбирается случайно, но с учетом пары **верных бит *b***

№ таблицы→		1	2	3	4	5	6	7	8	9	10	11	12
ϕ(y)	0	«11»	«11»	«11»	«11»	«11»	«11»	«00»	«00»	«00»	«00»	«00»	«00»
	1	«00»	«00»	«01»	«01»	«10»	«10»	«01»	«01»	«10»	«10»	«11»	«11»
	2	«01»	«10»	«00»	«10»	«00»	«01»	«11»	«10»	«01»	«11»	«10»	«01»
	3	«10»	«01»	«10»	«00»	«01»	«00»	«10»	«11»	«11»	«01»	«01»	«10»

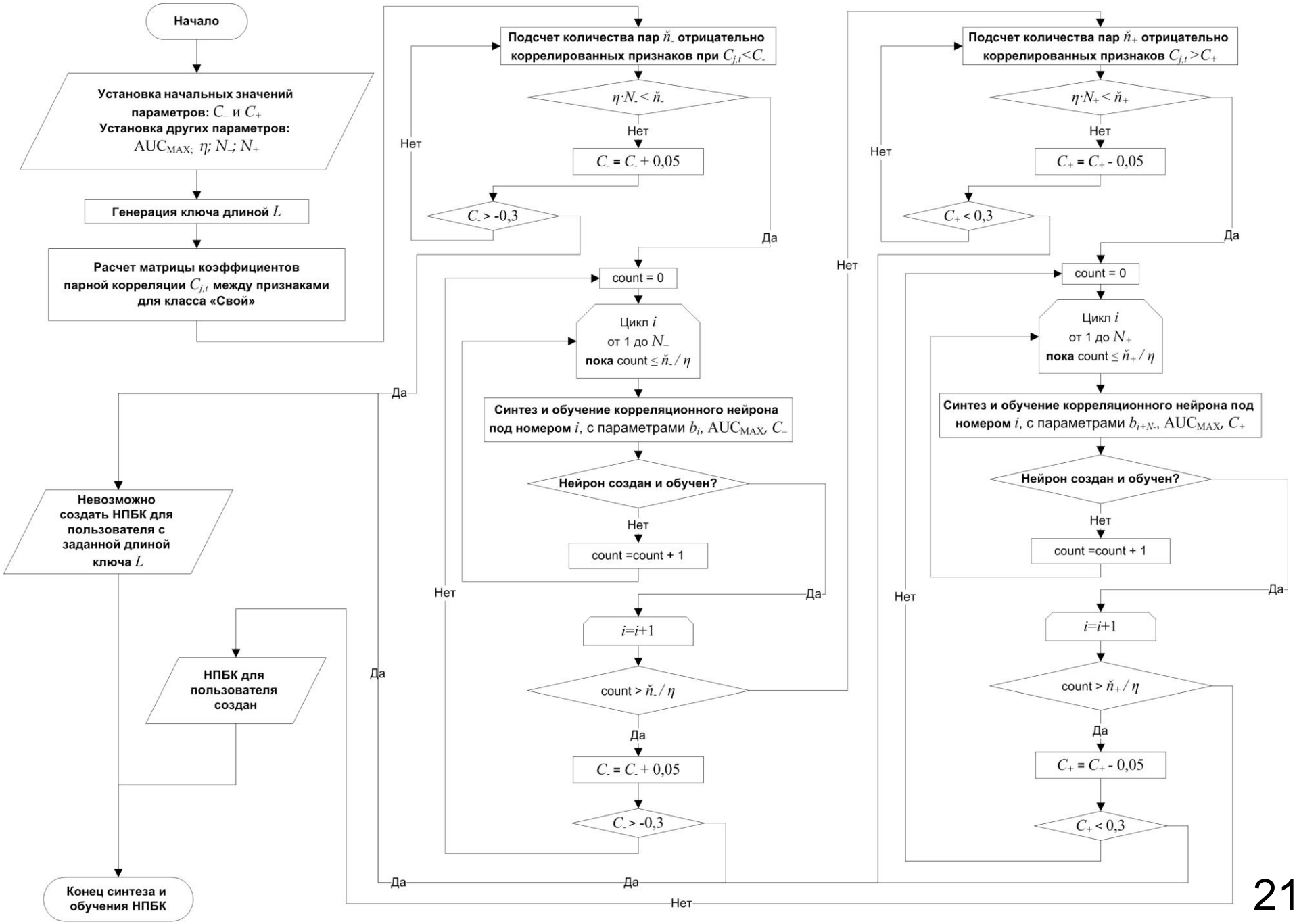
№ таблицы→		13	14	15	16	17	18	19	20	21	22	23	24
ϕ(y)	0	«01»	«01»	«01»	«01»	«01»	«01»	«10»	«10»	«10»	«10»	«10»	«10»
	1	«00»	«00»	«10»	«10»	«11»	«11»	«00»	«00»	«01»	«01»	«11»	«11»
	2	«11»	«10»	«00»	«11»	«10»	«00»	«01»	«11»	«11»	«00»	«00»	«01»
	3	«10»	«11»	«11»	«00»	«00»	«10»	«11»	«01»	«00»	«11»	«01»	«00»



Алгоритм автоматического синтеза и обучения НПБК

Преимущества предложенной модели НПБК и алгоритма обучения:

- полностью **автоматическое обучение** на небольших выборках
- потенциальное увеличение **длины ключа** и его энтропии за счет того, что **мета-признаков гораздо больше**
- потенциальное **снижение FRR и FAR** для слабых биометрических образов **из-за работы с корреляцией**, а не признаками
- атаки на ГОСТ Р 52633.5 более не применимы, так как **нейроны продуцируют более одного бита** ключа



Суть концепции защищенного исполнения нейросетевых алгоритмов ИИ