# Boosting Novelty Detection Neural Networks with Rational Activations

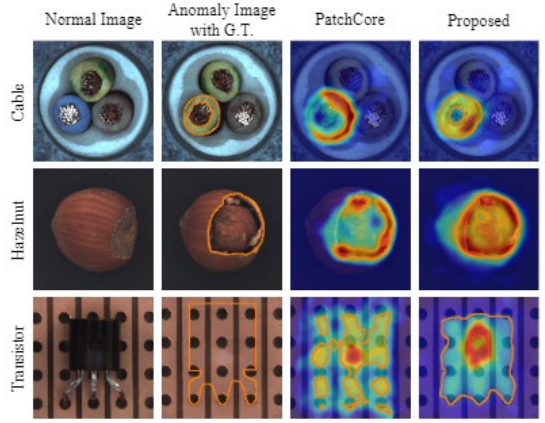
Zaborenko A.(1,2), Abasov E.(1,2), Boos E.(1), Bunichev V.(1), Volkov P.(1), Vorotnikov G.(1), Dudko L.(1), Iudin E.(1,2), Markina A.(1), Perfilov M.(1)

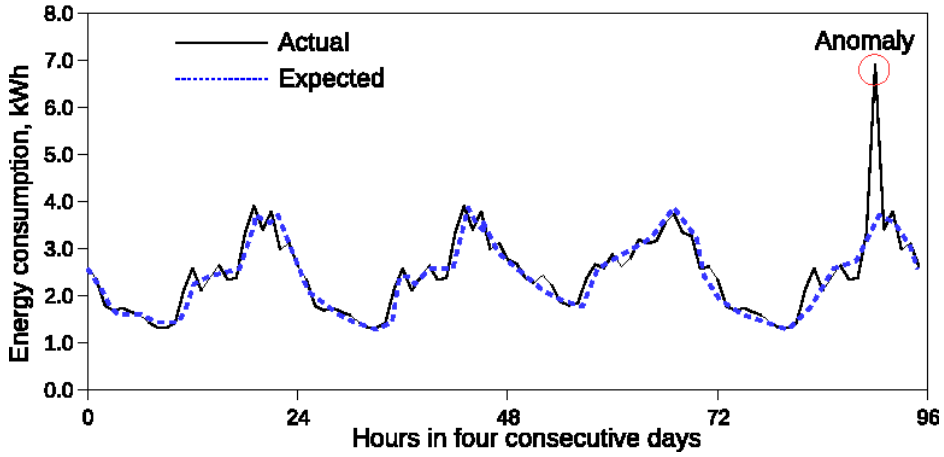
(1) Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Leninskie gory, GSP-1, Moscow 119991, Russian Federation, (2) Lomonosov Moscow State University, Faculty of Physics, Leninskie gory, GSP-1, Moscow 119991,Russian Federation

This report was conducted within the scientific program of the Russian National Center for Physics and Mathematics, section 5 ≪Particle Physics and Cosmology≫.

# Детектирование аномалий

- Детектирование аномалий – процесс поиска отклонений от данных, считающихся «нормальными» без знаний о конкретной природе таких отклонений.
- Задача трудна: имея знания только об одном классе, нужно понять, что считать «отличием», а что – вариацией нормального класса. Нужно «провести» бинарную классификацию без разметки второго класса.

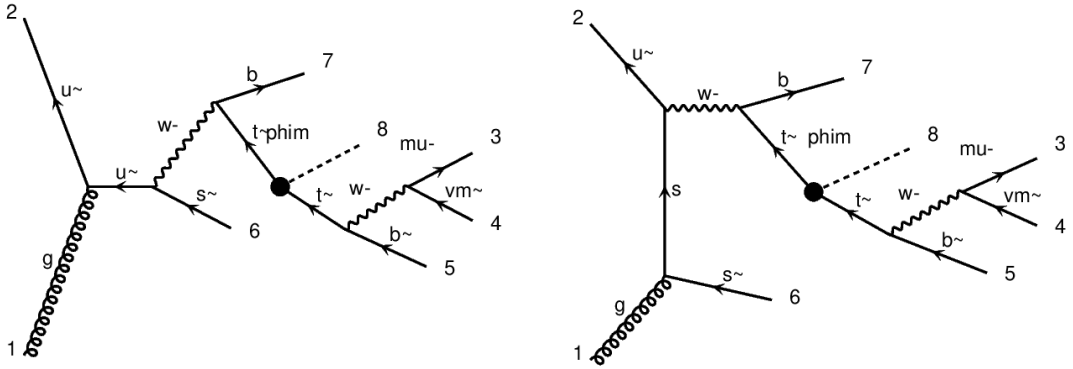



Детектирование аномалий в индустрии

# Детектирование аномалий: HEP

- Задача поиска аномалий в HEP: произвести **модельно-независимый** отбор событий.*
- Алгоритм поиска аномалий обучается на событиях Стандартной модели (СМ) и детектирует значимые отклонения от СМ в данных.
- Эффективность алгоритмов была оценена в задаче разделения событий Стандартной Модели от событий ассоциированного рождения топ-кварка с медиатором Темной Материи (ТМ).

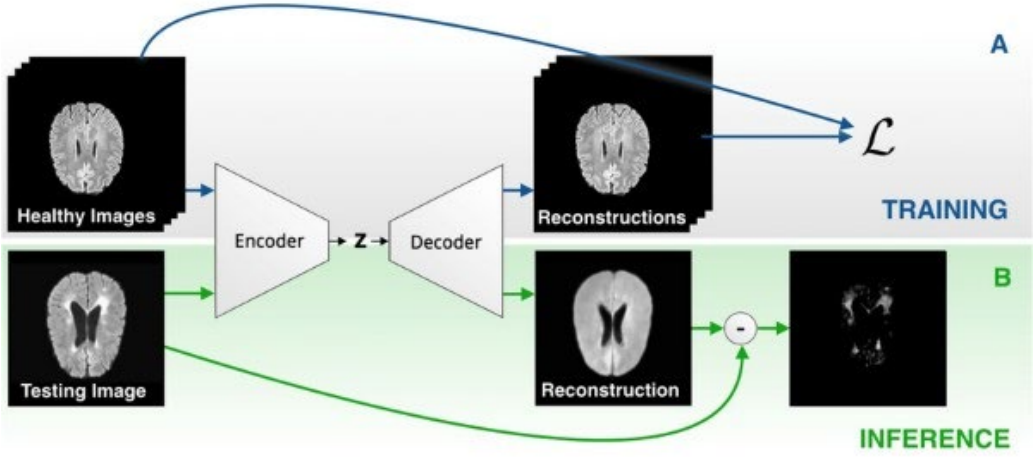



Фейнмановские диаграммы характерных событий с присутствием медиатора Темной Материи.

*Более детально классический подход описан в E. E. Abasov, M. I. Belobrova, P. V. Volkov, G. A. Vorotnikov, L. V. Dudko, A. D. Zaborenko, M. A. Perfilov и E. S. Sivakova. «Methodology for the Application of Deep Neural Networks in Searches for New Physics at Colliders and Statistical Interpretation of Expected Results». В: Phys. Atom. Nucl.85.6 (2022), с. 708—720. doi: 10.1134/S1063778822060023.

# Классический подход: автоэнкодер

- Автоэнкодер – это нейронная сеть с «бутылочным горлышком»: слоем с числом нейронов меньшим, чем размерность входных данных.
- Задача такой нейронной сети – максимально точно реконструировать входные данные, пропустив через «бутылочное горлышко» только самые важные компоненты.

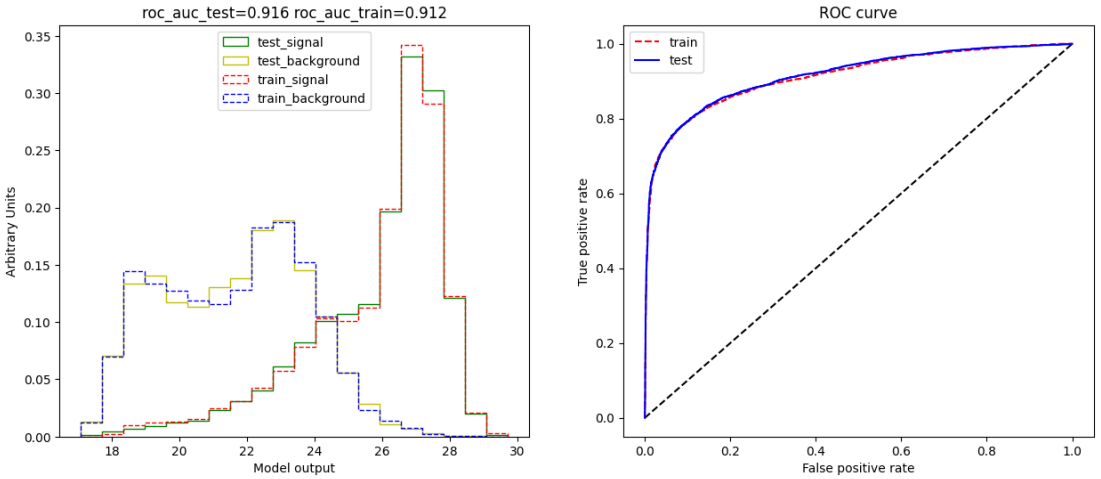



Для детектирования аномалий используется ошибка реконструкции: сеть будет хуже реконструировать те данные, которые она не «видела» при обучении.

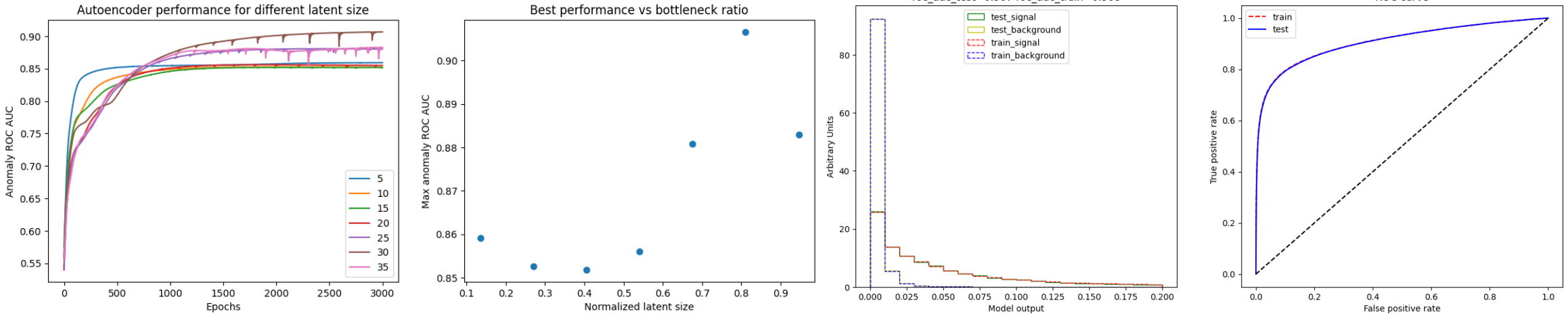
# Автоэнкодер в HEP

- Для данных физики высоких энергий автоэнкодер показал себя недостаточно эффективным: простой классификатор по принципу ближайших соседей давал лучшие результаты.

KNN

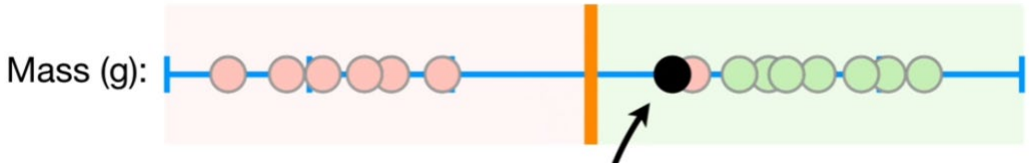



Autoencoder, 30 latent units

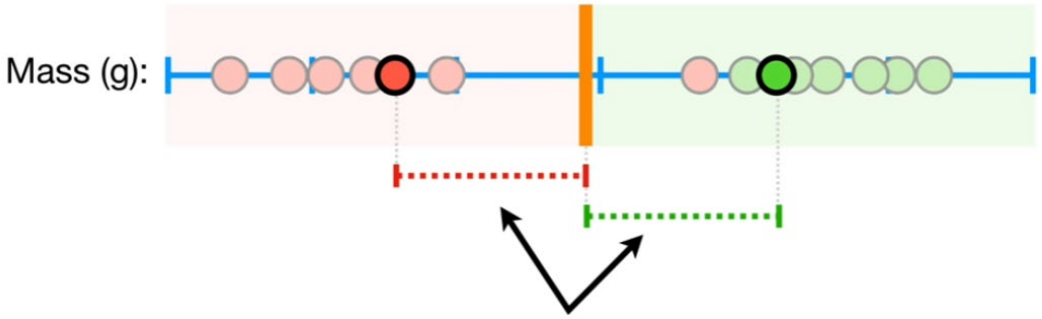


# Метод Опорных Векторов (SVM)

- Метод опорных векторов:
  - Отображает данные в пространство более высокой размерности, используя специальную функцию – «ядро»
  - Разделяет классы с «правом на ошибку»: позволяет работать с шумными данными, некоторая часть точек может попасть в противоположный класс
  - Использует кросс-валидацию для поиска оптимальной классификации



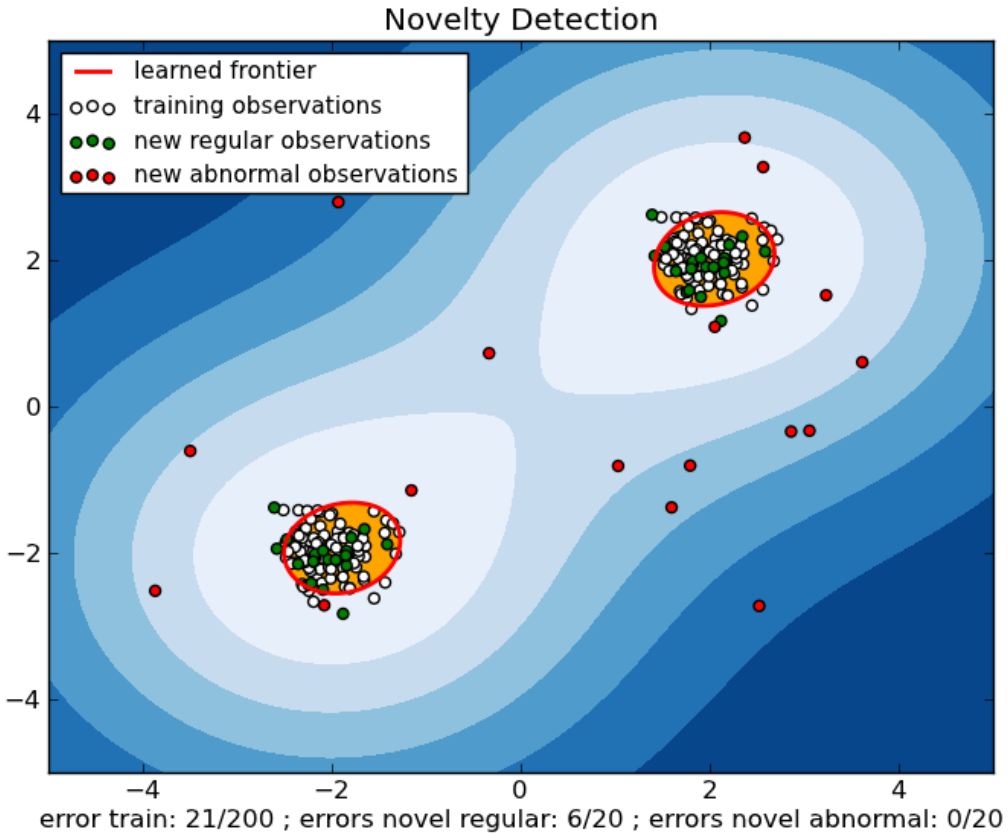

Установление порога с «правом на ошибку» с помощью кросс-валидации




Опорные векторы

# Метод Опорных Векторов для одного класса

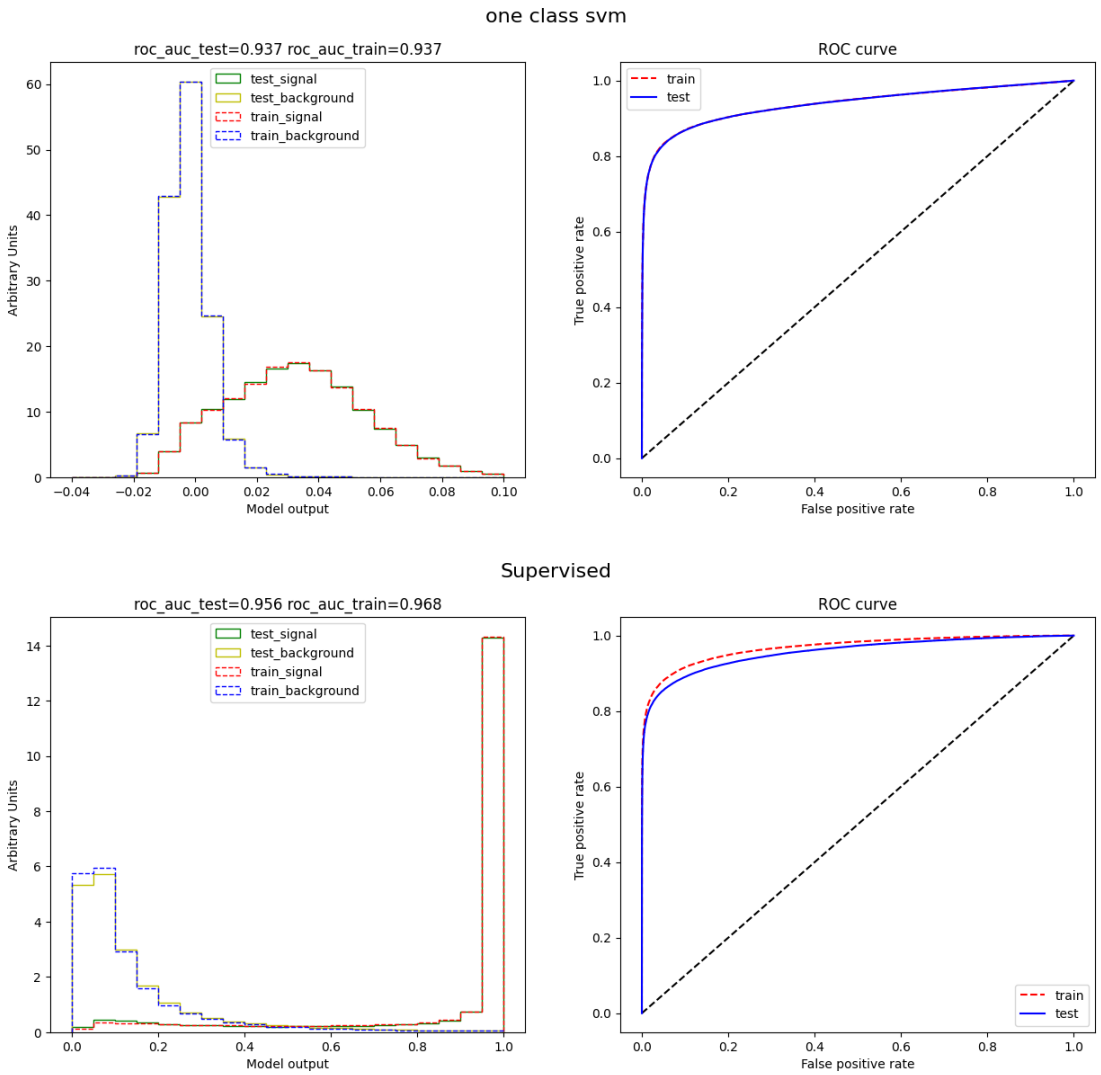
- SVM может создавать гиперплоскость вокруг данных, позволяя выделять аномалии как выход за эту гиперплоскость.
- С правильными гиперпараметрами может быть очень точен.
- Минус – время растет квадратично/кубично с числом примеров.




Support Vector Machines are powerful tools, but their compute and storage requirements increase rapidly with the number of training vectors. The core of an SVM is a quadratic programming problem (QP), separating support vectors from the rest of the training data. The QP solver used by the libsvm-based implementation scales between $O(n_{features} \times n_{samples}^2)$ and $O(n_{features} \times n_{samples}^3)$ depending on how efficiently the libsvm cache is used in practice (dataset dependent). If the data is very sparse $n_{features}$ should be replaced by the average number of non-zero features in a sample vector.

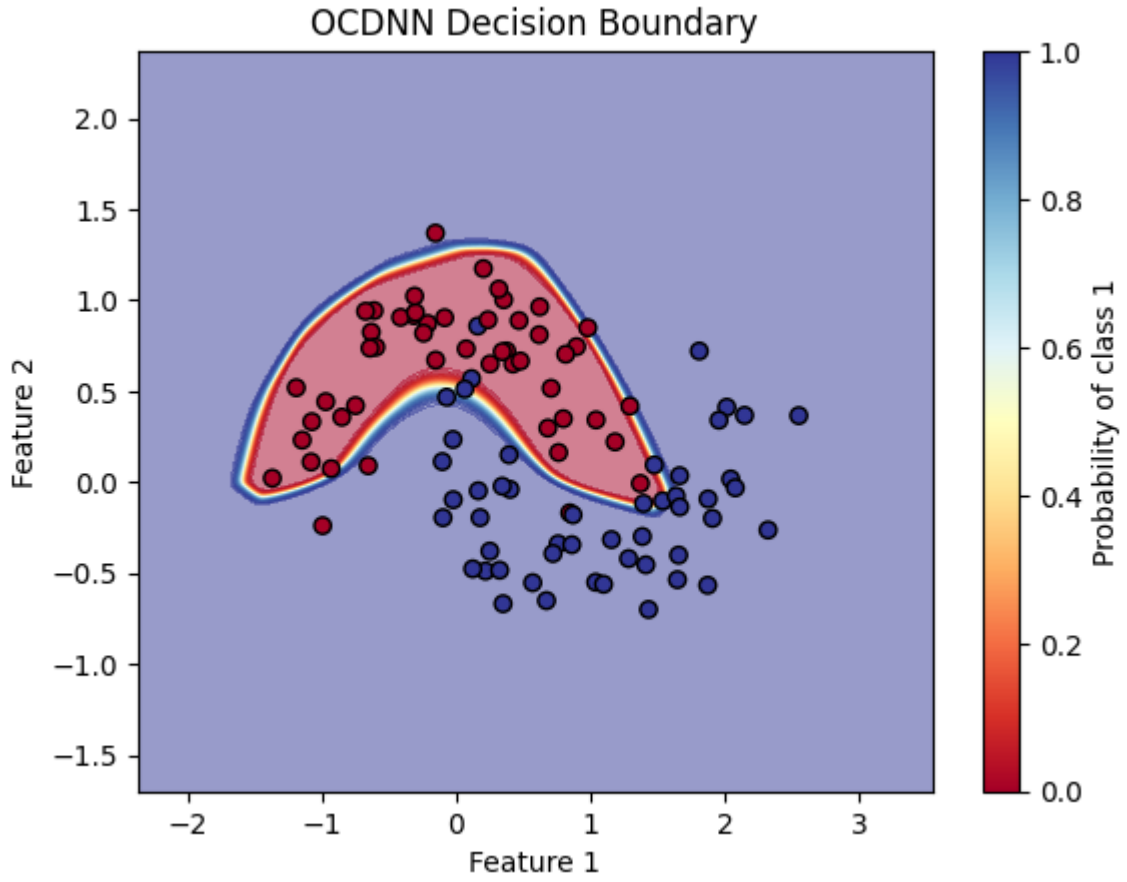
# SVM: данные HEP

- С правильными параметрами позволяет приблизиться к обучению с учителем.
- Низкая скорость предсказаний ограничивает использование в физике, где обрабатываются миллионы событий.



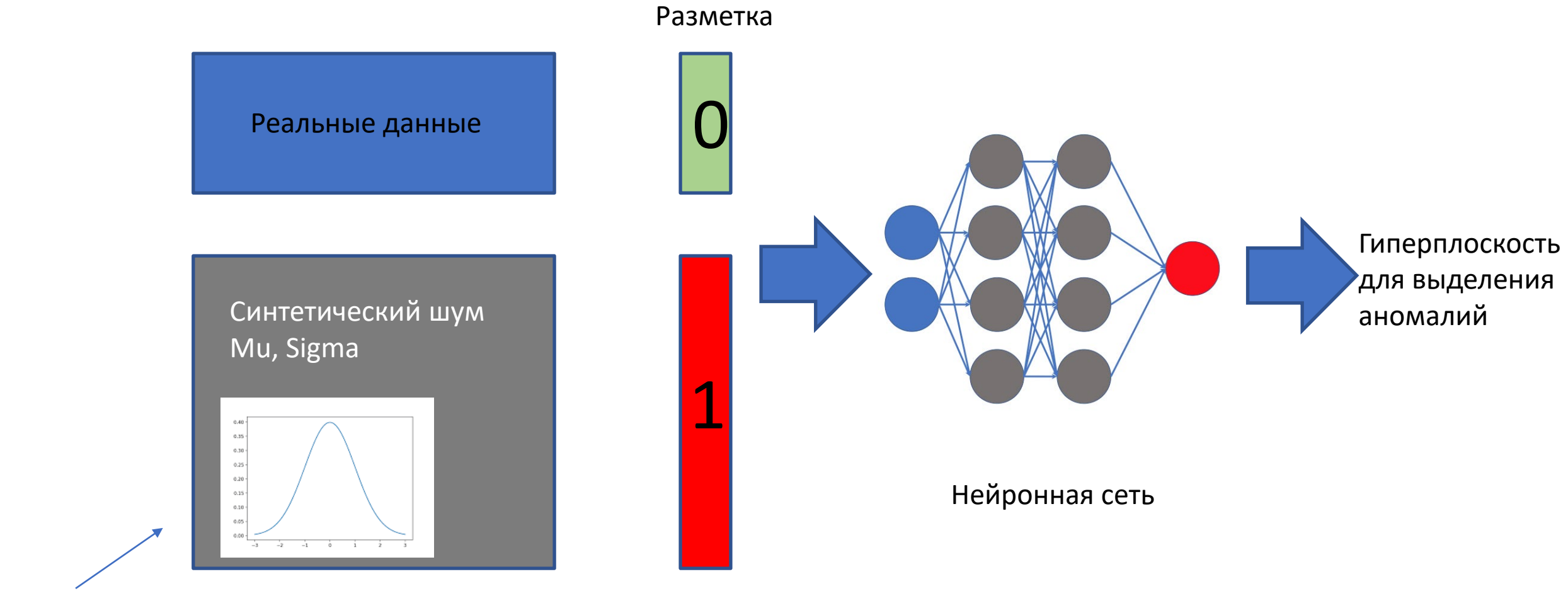
# Нейронная сеть для одного класса

- Основная концепция:
  - генерировать синтетический шум* каждую эпоху
  - обучать сеть отделять шум от реальных данных
  - использовать выходные данные сети в качестве оценки аномалий
- Во время обучения сеть строит поверхность, окружающую данные, как алгоритм One Class SVM. (В задаче ТМ ранговая корреляция 0.98)



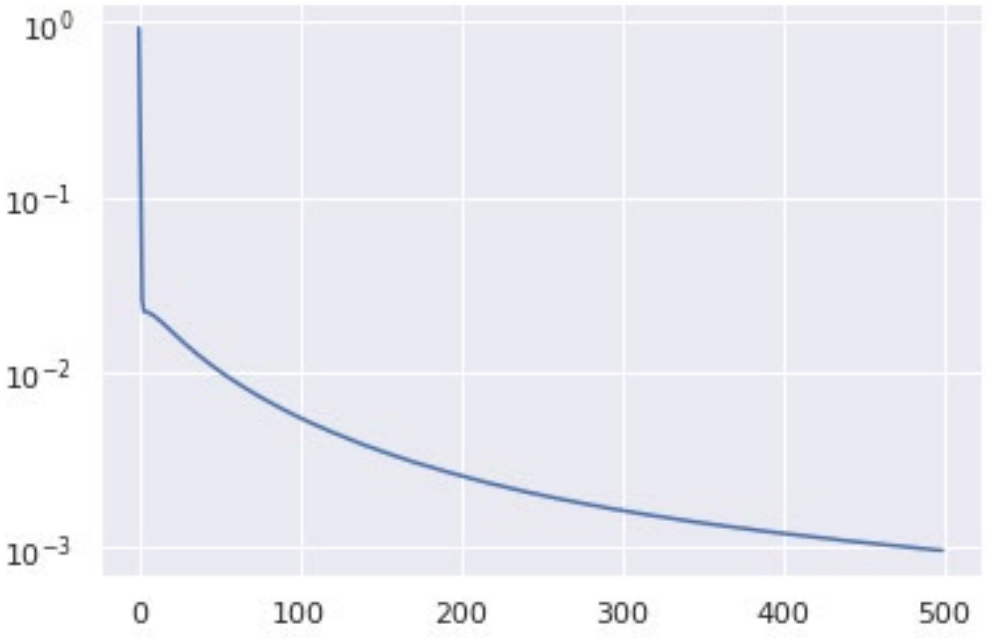

*Подход похож на метод, использованный в P. Oza and V. M. Patel, "One-Class Convolutional Neural Network," in *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277-281, Feb. 2019, doi: 10.1109/LSP.2018.2889273.

# Нейронная сеть для одного класса*: схема

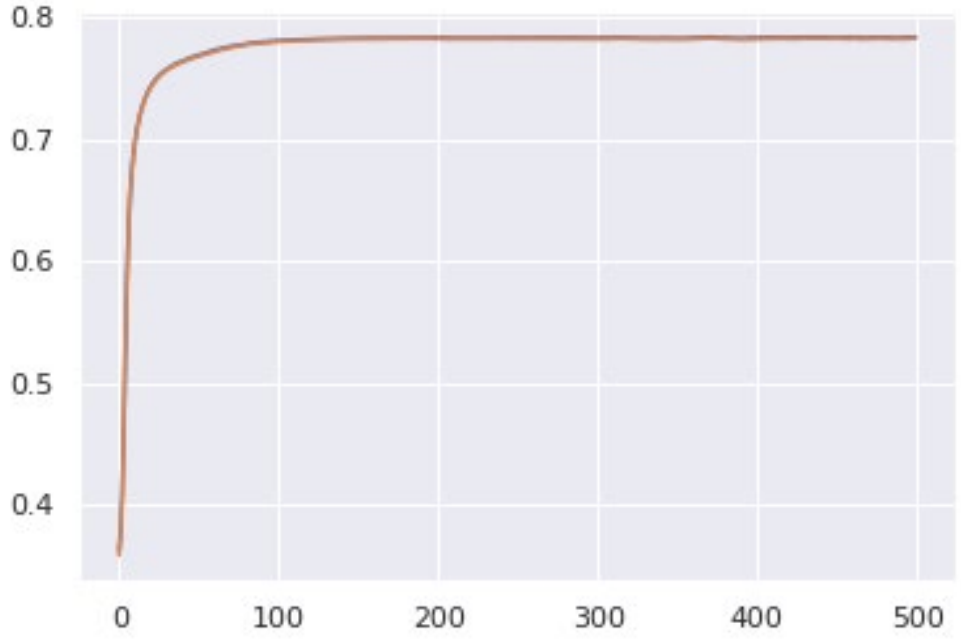



*ISSN 0027-1349, Moscow University Physics Bulletin, 2023, Vol. 78, No. 7, pp. 80–84.* **DOI:** 10.3103/S0027134923070329

# Нейронная сеть для одного класса: обучение




Функция ошибки для разделения нормального класса от шума*

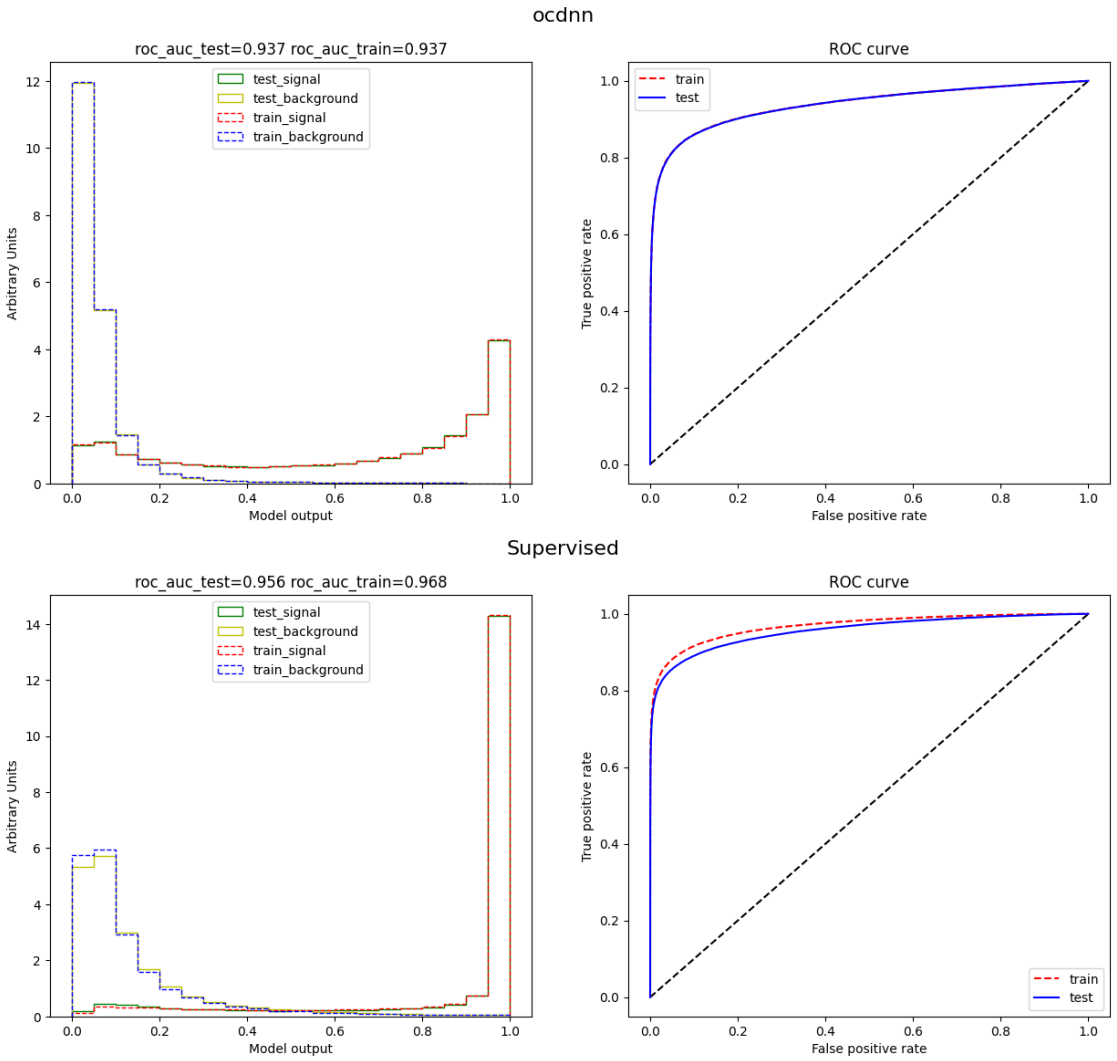



Способность сети выделять аномалии

*Unleashing the Potential of Unsupervised Deep Outlier Detection through Automated Training Stopping: Huang et al

# Нейронная сеть для одного класса: результаты

- Нейронная сеть может обучаться и делать предсказания гораздо быстрее, чем SVM (Скорость предсказаний выше в 14,000 раз)
- В более сложных задачах нейронная сеть выучивает более «тонкие» корелляции в данных, позволяя повысить точность детектирования аномалий.



# Упрощенные модели ТМ с участием топ-кварка

Одними из наиболее простых моделей темной материи являются так называемые «упрощенные модели», в которых предполагается, что частицы ТМ взаимодействуют с частицами СМ, обмениваясь одной или несколькими частицами, называемыми "медиаторами", которые обладают слабой связью с частицами СМ.

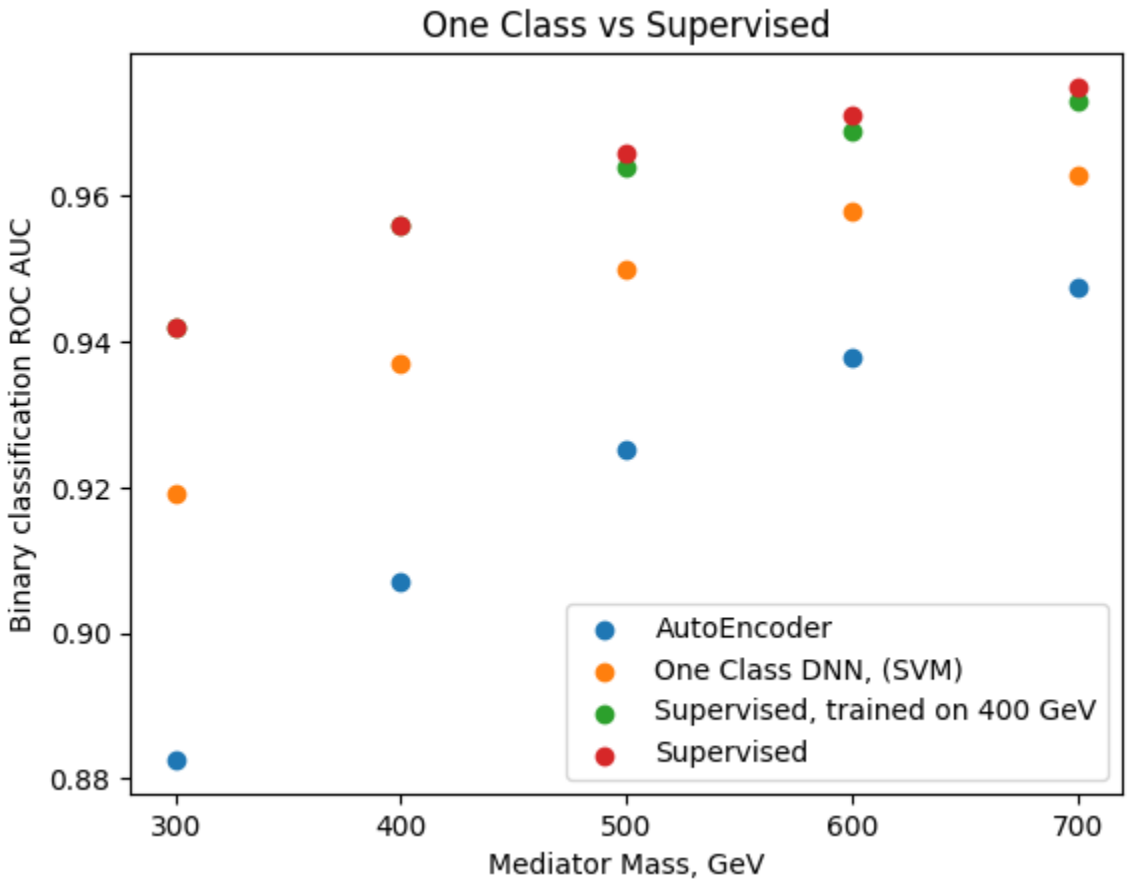
$$L_\Phi = g_\chi \Phi \bar{\chi}\chi + \frac{g_v \Phi}{\sqrt{2}} \sum_f \left( y_f \bar{f} f \right)$$

$$L_A = i g_\chi A \bar{\chi}\gamma^5 \chi + i \frac{g_v A}{\sqrt{2}} \sum_f \left( y_f \bar{f} \gamma^5 f \right)$$

Это приводит к возможности производства частиц СМ вместе с частицами ТМ и, соответственно, к наблюдению характерной "потерянной энергии" в таких процессах, поскольку частицы ТМ не обнаруживаются напрямую. Совместное производство частиц СМ и ТМ с последующим обнаружением потерянной энергии – единственный способ обнаружить данные процессы на коллайдерах.

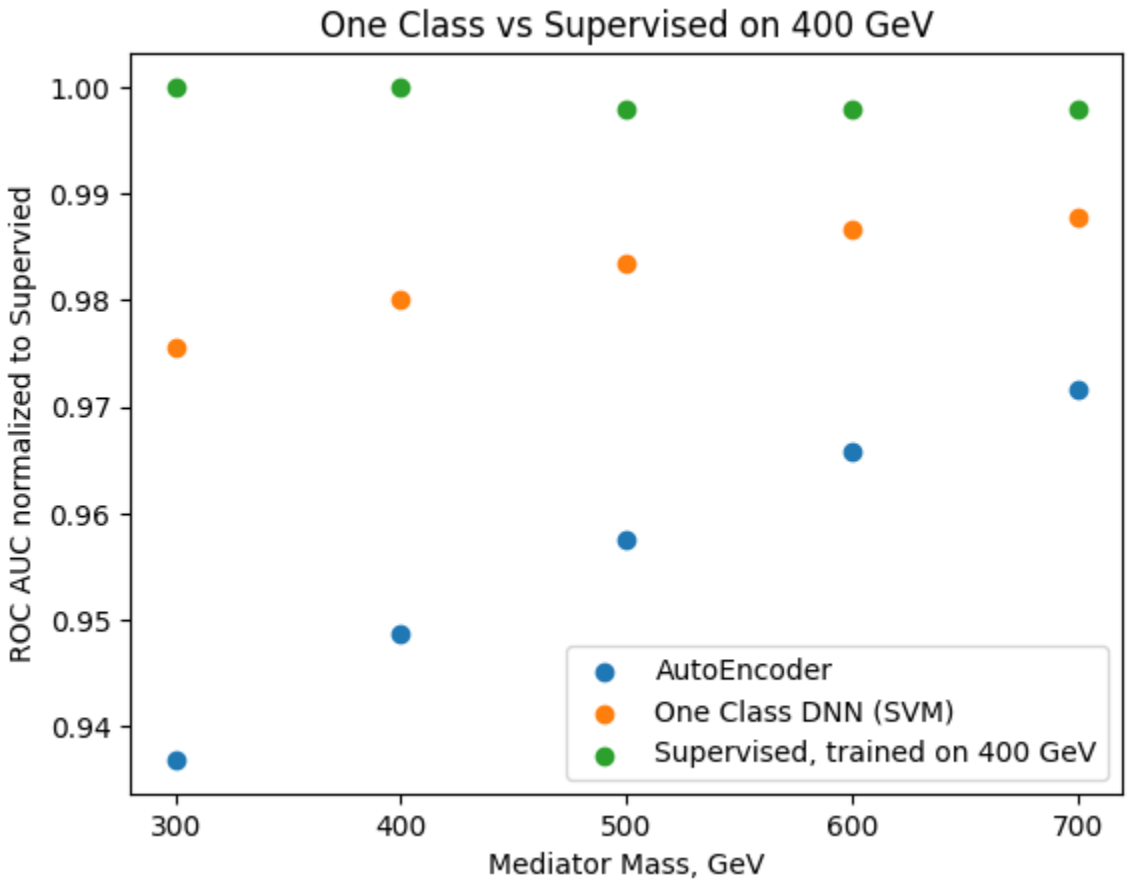
# Применение алгоритмов детектирования аномалий к событиям упрощенной модели ТМ

- Вышеописанные алгоритмы были протестированы на событиях, сгенерированных с разной массой скалярного медиатора ТМ. Методы, обученные только на СМ, были противопоставлены классификатору, обученному «с учителем» на событиях ТМ с медиатором с массой 400 ГэВ.



# Применение алгоритмов детектирования аномалий к событиям упрощенной модели ТМ

- На приведенном графике показаны метрики ROC AUC алгоритмов детектирования аномалий, нормированные на метрику классификатора «с учителем» на данных с указанной массой медиатора.
- OCDNN и OCSVM показывают отличную точность классификации, их метрика лежит в 97-98% от «идеального» классификатора.
- Чем сильнее отличаются данные от обучающей выборки классификатора, обученного на массе 400 ГэВ, тем ниже его относительная классификационная способность, что является оптимистичным прогнозом для ОС методов, не зависящих от конкретной сигнатуры сигнального процесса.



# Дальнейшее улучшение алгоритма

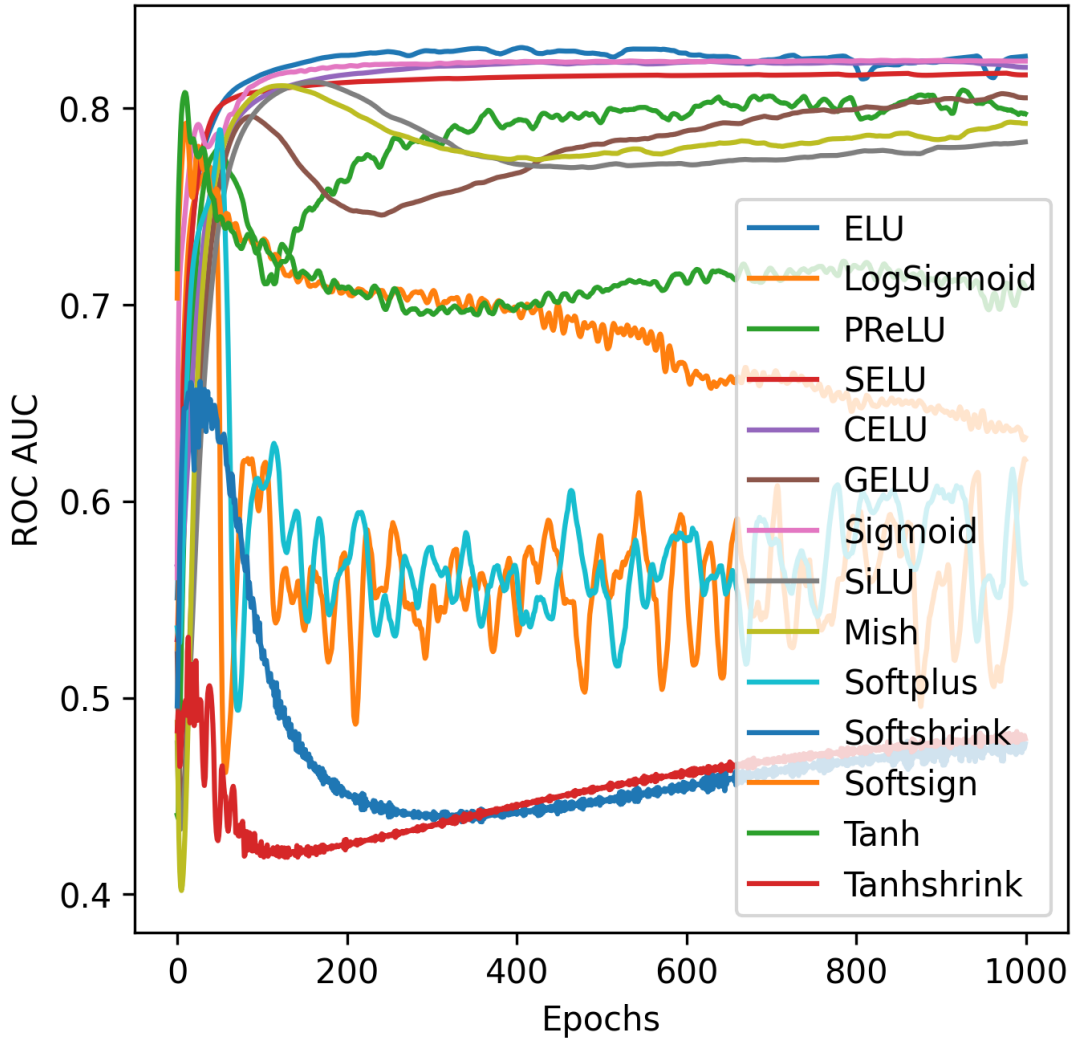
- Алгоритм OCDNN был эффективно применен к реальным задачам HEP, показал свою точность в сравнении с классическими и новыми «глубокими» алгоритмами детектирования аномалий.
- Предварительные результаты в сравнении с алгоритмами из пакета DeepOD:

| Model | Venue | Year | Type | Title |
|---|---|---|---|---|
| Deep SVDD | ICML | 2018 | unsupervised | Deep One-Class Classification [1] |
| REPEN | KDD | 2018 | unsupervised | Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection [2] |
| RDP | IJCAI | 2020 | unsupervised | Unsupervised Representation Learning by Predicting Random Distances [3] |
| RCA | IJCAI | 2021 | unsupervised | RCA: A Deep Collaborative Autoencoder Approach for Anomaly Detection [4] |
| GOAD | ICLR | 2020 | unsupervised | Classification-Based Anomaly Detection for General Data [5] |
| NeuTraL | ICML | 2021 | unsupervised | Neural Transformation Learning for Deep Anomaly Detection Beyond Images [6] |
| ICL | ICLR | 2022 | unsupervised | Anomaly Detection for Tabular Data with Internal Contrastive Learning [7] |
| DIF | TKDE | 2023 | unsupervised | Deep Isolation Forest for Anomaly Detection [19] |
| SLAD | ICML | 2023 | unsupervised | Fascinating Supervisory Signals and Where to Find Them: Deep Anomaly Detection with Scale Learning [20] |

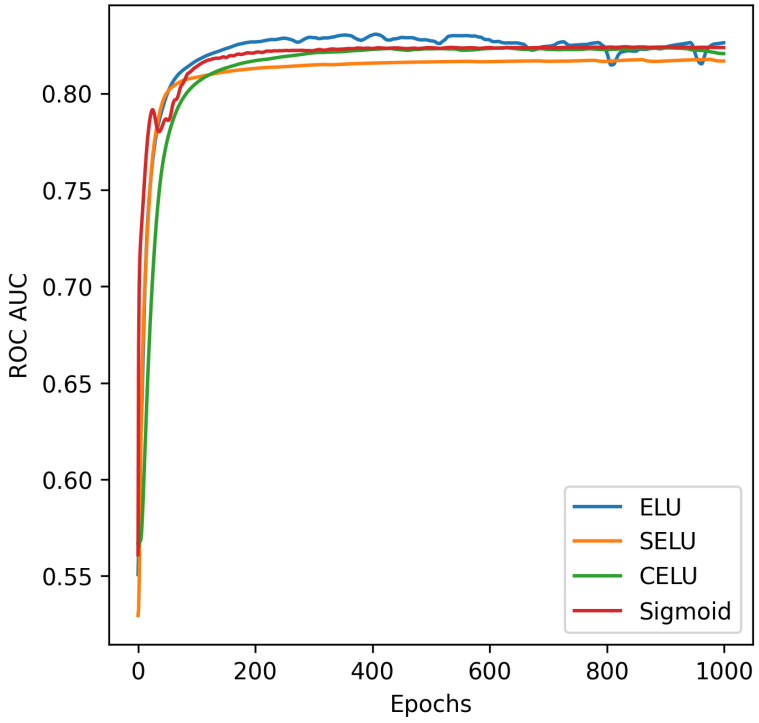
| Algorithm | OCDNN | OCSVM | NeuTral | SLAD | GOAD | ICL | RDP |
|---|---|---|---|---|---|---|---|
| Normalized AUC | 0.97 | 0.95 | 0.84 | 0.74 | 0.73 | 0.66 | 0.65 |

# Дальнейшее улучшение алгоритма

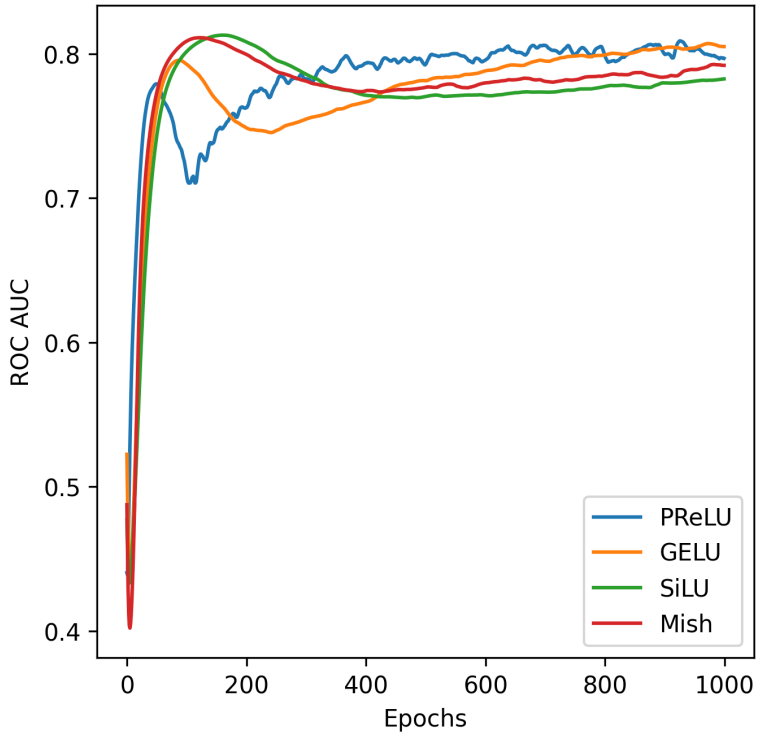
- Было обнаружено, что выбор функции активации значительно влияет на точность финального алгоритма и стабильность тренировки.
- Можно выделить три глобальные группы функций активации, и на одной мы заострим внимание.



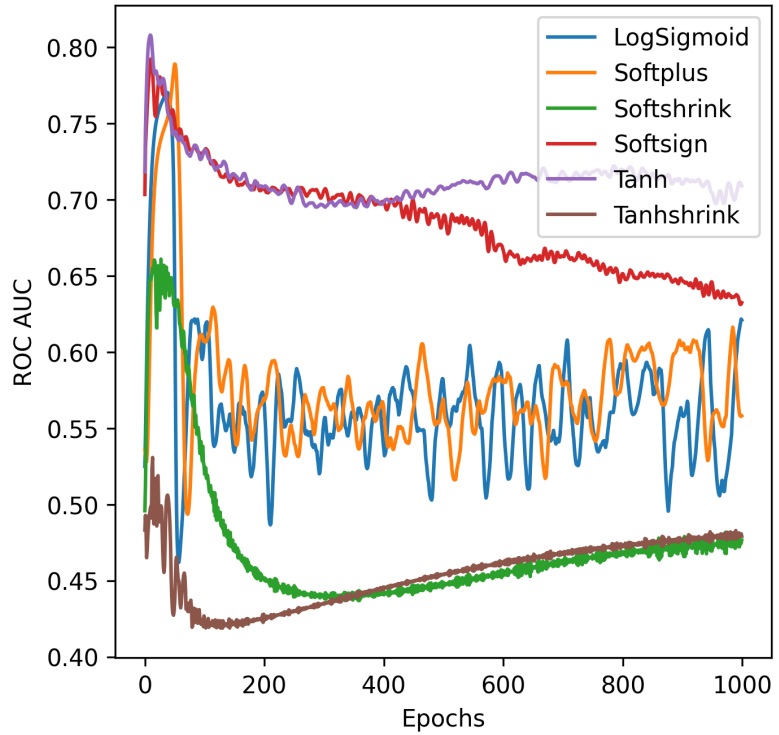
# Функции активации OCDNN




Экспоненциальные



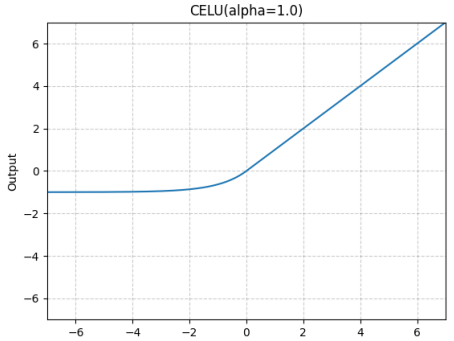

Современные ReLU-based




Неподходящие

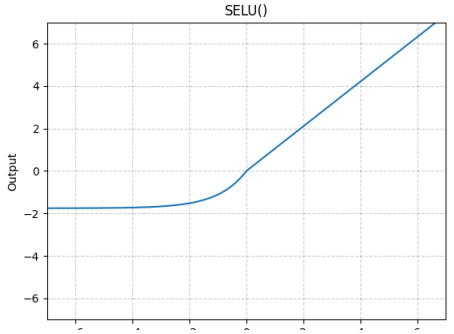
# Экспоненциальные функции активации

$$\mathrm{CELU}(x) = \max(0, x) + \min(0, \alpha * (\exp(x/\alpha) - 1))$$

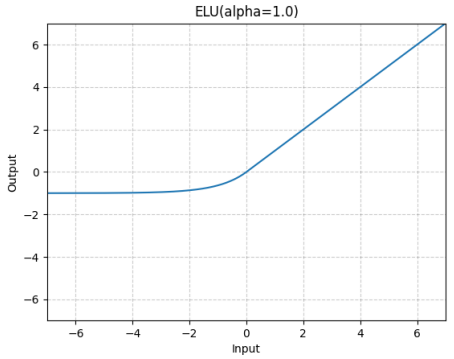


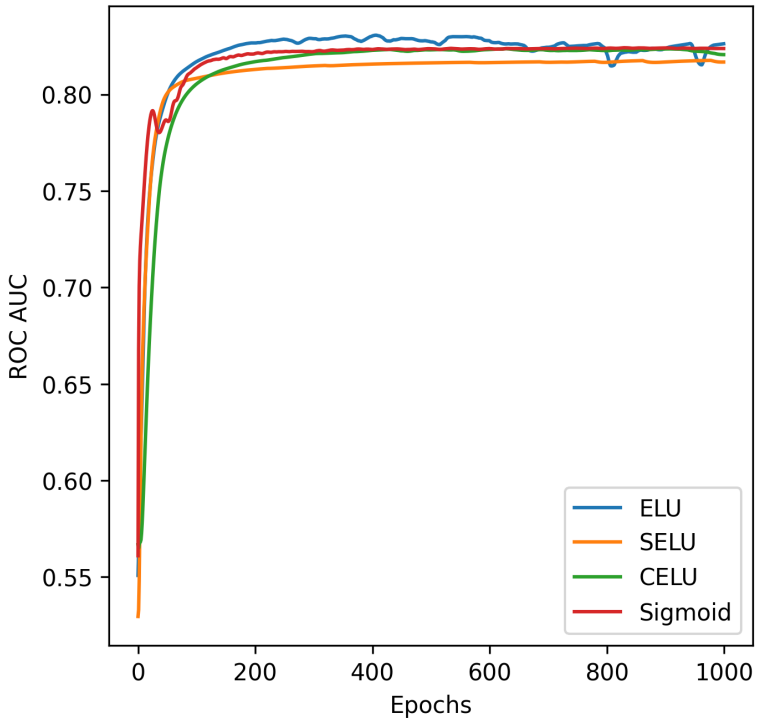

$$\mathrm{SELU}(x) = \mathrm{scale} * (\max(0, x) + \min(0, \alpha * (\exp(x) - 1)))$$

with $\alpha = 1.6732632423543772848170429916717$ and $\mathrm{scale} = 1.0507009873554804934193349852946$.




$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha * (\exp(x) - 1), & \text{if } x \leq 0 \end{cases}$$

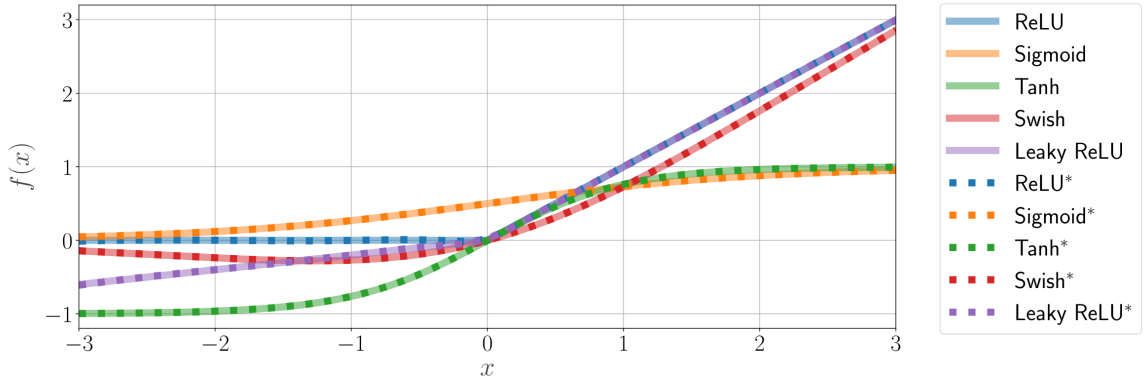


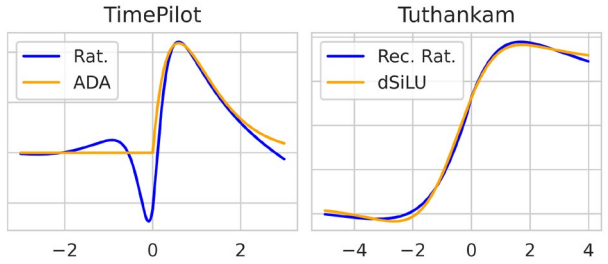



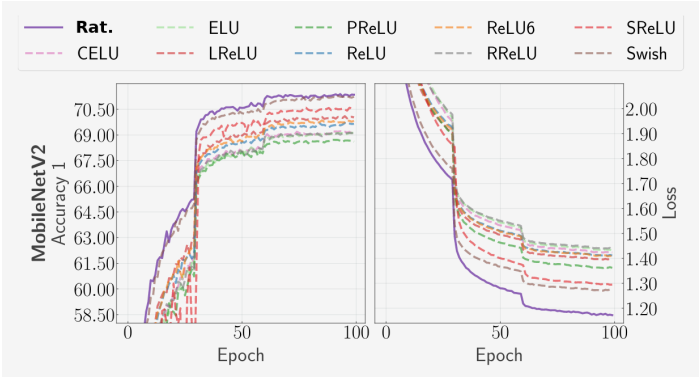

Экспоненциальные функции активации вели себя стабильно и достигали максимальных результатов

# Поиск функций активации

- Так как выбор более оптимальной функции активации позволяет повысить эффективность сети, возникла идея использовать обучаемые функции активации.
- Одной из таких функций активации является Рациональная активация (Rational Activation).




https://doi.org/10.48550/arXiv.2004.01902

# Алгоритм рациональной активации

- Функция активации задается отношением полиномов, коэффициенты которых выучиваются с помощью обратного распространения ошибки.
- На каждом слое можно задать свою функцию активации, которая будет меняться в процессе обучения:

$$F(x) = \frac{P(x)}{Q(x)} = \frac{\sum_{j=0}^{m} a_j x^j}{1 + |\sum_{k=1}^{n} b_k x^k|} = \frac{a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m}{1 + |b_1 x + b_2 x^2 + \cdots + b_n x^n|}.$$

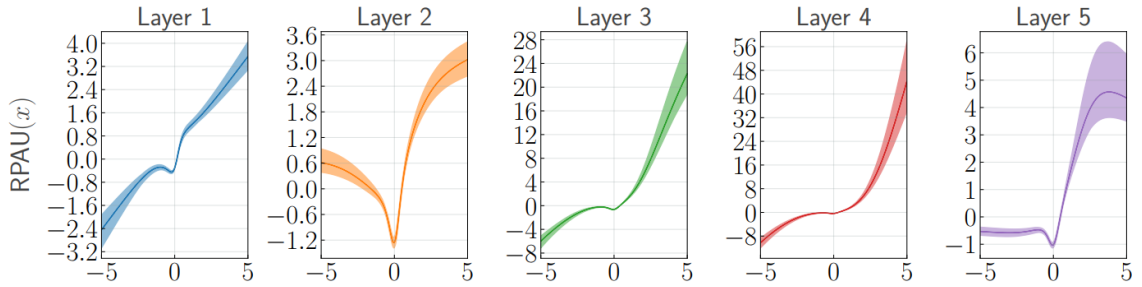
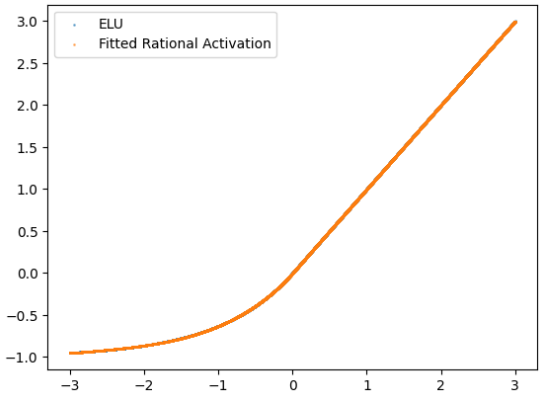


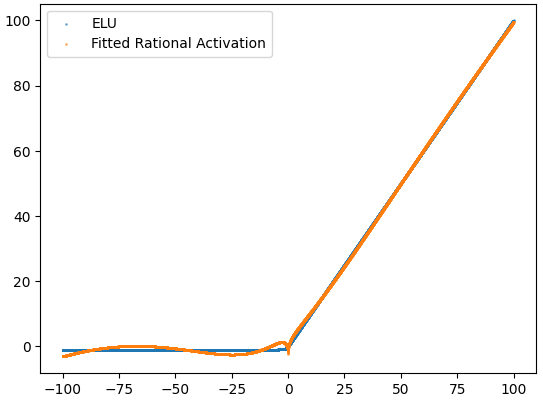

Figure 3: Estimated activation functions after training the VGG-8 network with RPAU on Fashion-MNIST. The center line indicates the PAU while the surrounding area indicates the space of the additive noise in RPAUs. As one can see, the PAU family differs from common activation functions but capture characteristics of them. (Best viewed in color)

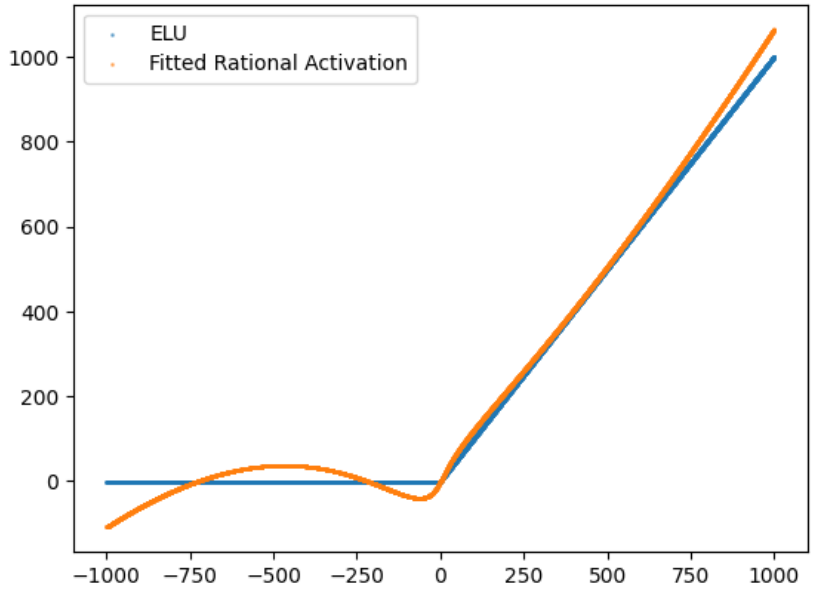
https://doi.org/10.48550/arXiv.1907.06732

# Применение к экспоненциальным функциям

- Авторы статьи используют аппроксимацию известных функций активации для начальной инициализации коэффициентов полиномов.
- В статье используется аппроксимация на отрезке [-3, 3], на котором экспоненциальная функция хорошо фитируется.
- Несмотря на это, пока не удалось добиться хорошей сходимости сети с использованием такой отфитированной функции, даже с «замороженными» весами. Возможно, это связано с поведением функции активации на большом отдалении от нуля.



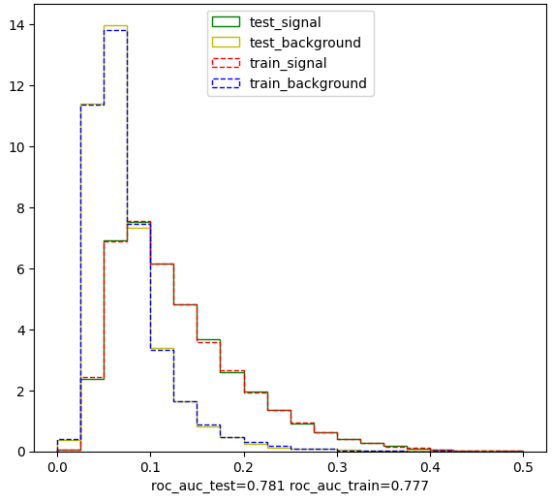



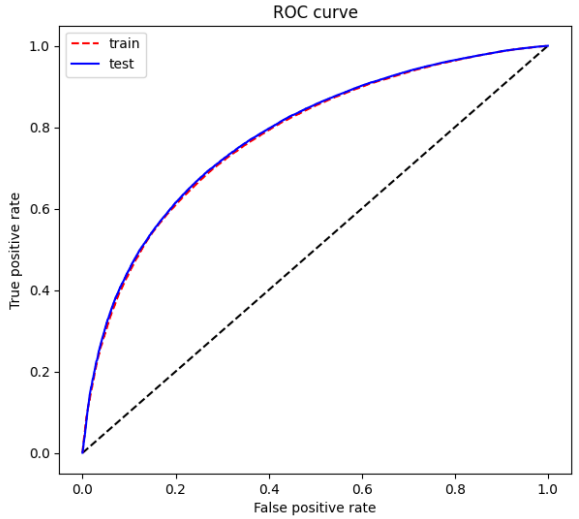



# Краткие итоги и перспективы

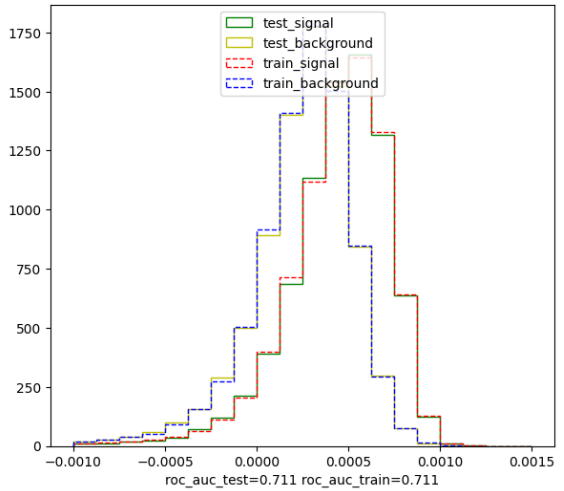
- Была продемонстрирована эффективность алгоритма OCDNN на реальных задачах Физики Высоких Энергий: в «простых» задачах работает с выскокой точностью SVM, но гораздо быстрее; в «сложных» задачах работает лучше за счет многослойной структуры.
- Были представлены исследования зависимости точности алгоритма от функции активации: перспективная группа экспоненциальных функций, ведется работа по их параметризации с помощью обучаемых функций активации.

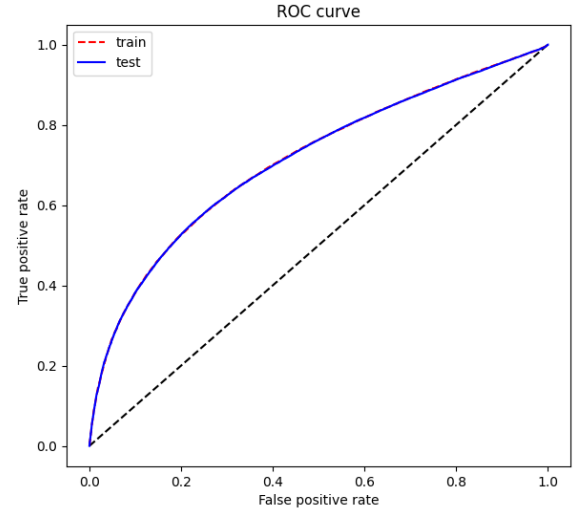
ocdnn, t-channel




one class svm, t-channel



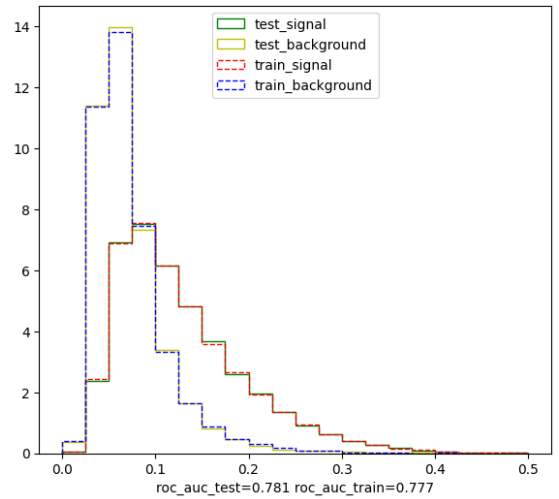



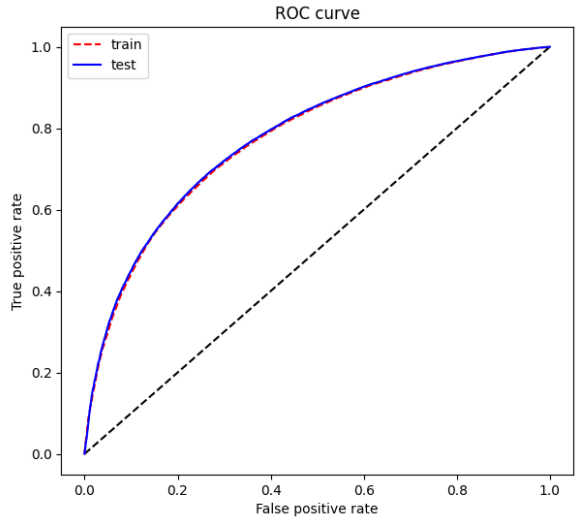
Спасибо за внимание!

# One Class SVM vs One Class DNN

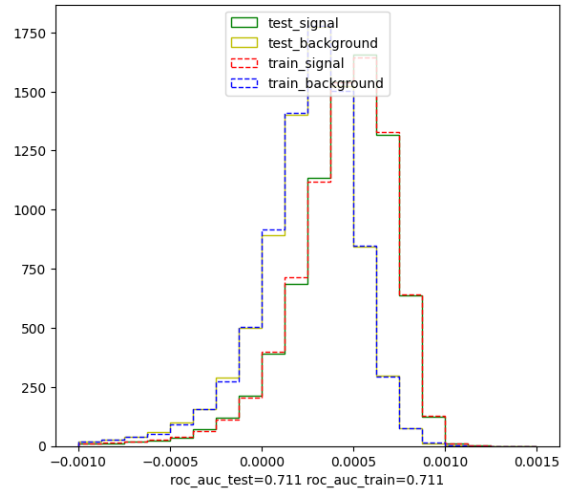
- На более сложных датасетах многослойная структура нейронной сети позволяет ей выделить более сложные корелляции в данных, чем это может сделать SVM.
- Представлено сравнение алгоритмов в задаче выделения т-канального рождения топ-кварка из СМ фона.

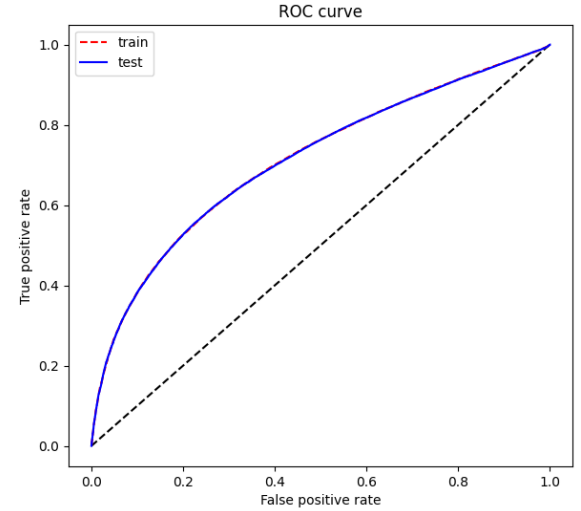
ocdnn, t-channel



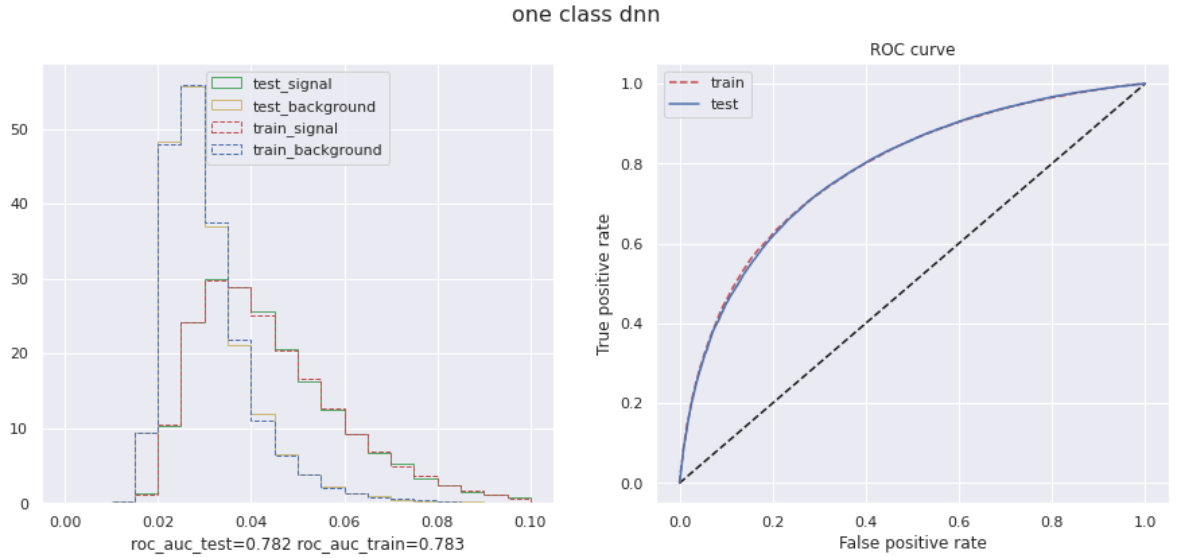

one class svm, t-channel

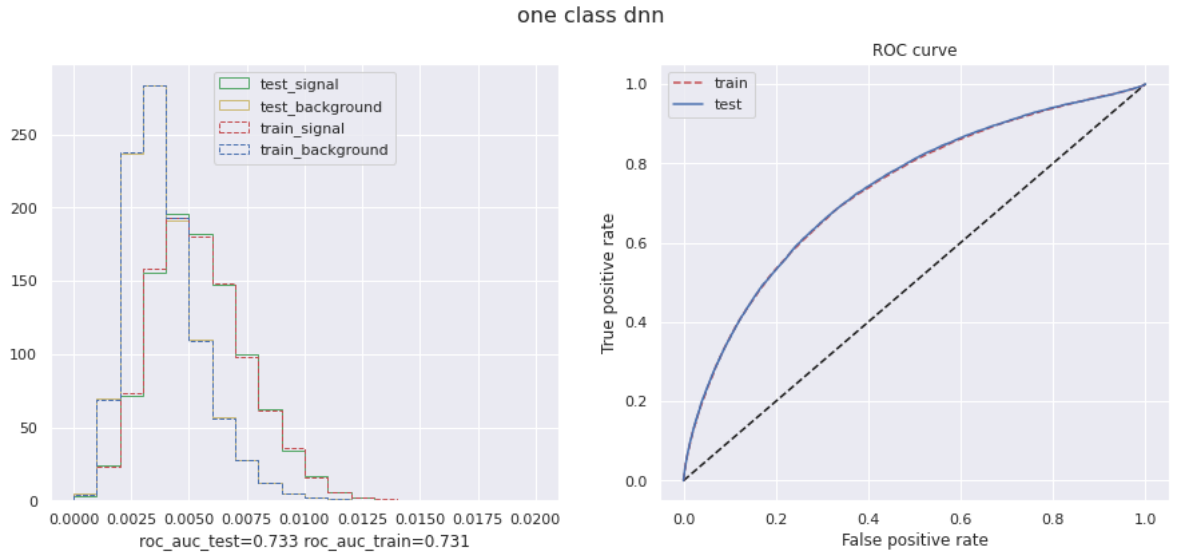


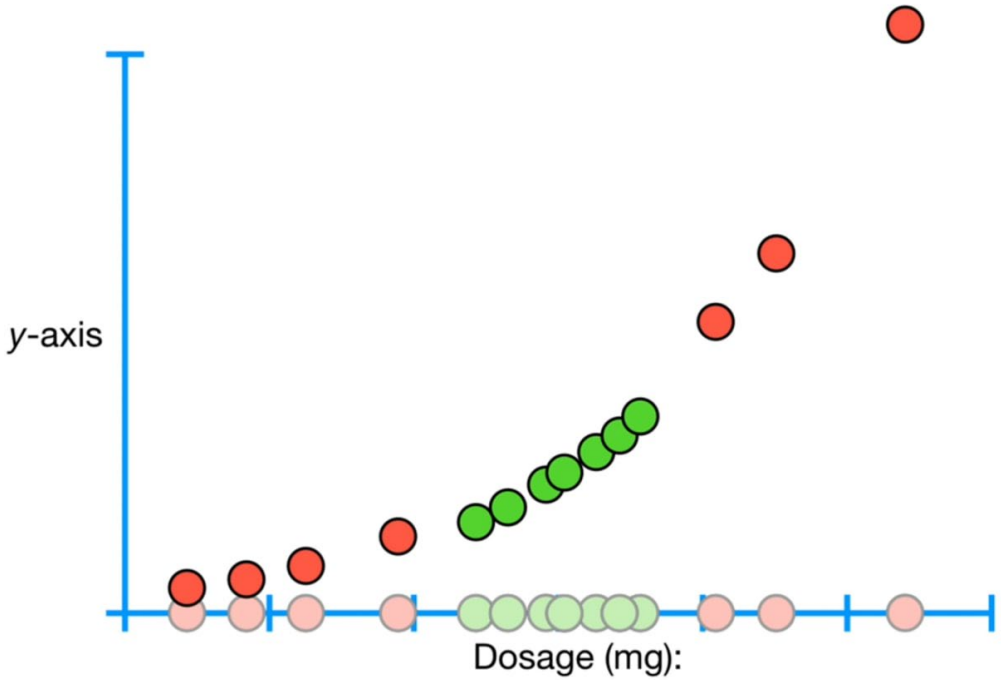
# Нейронная сеть для одного класса: реальные датасеты

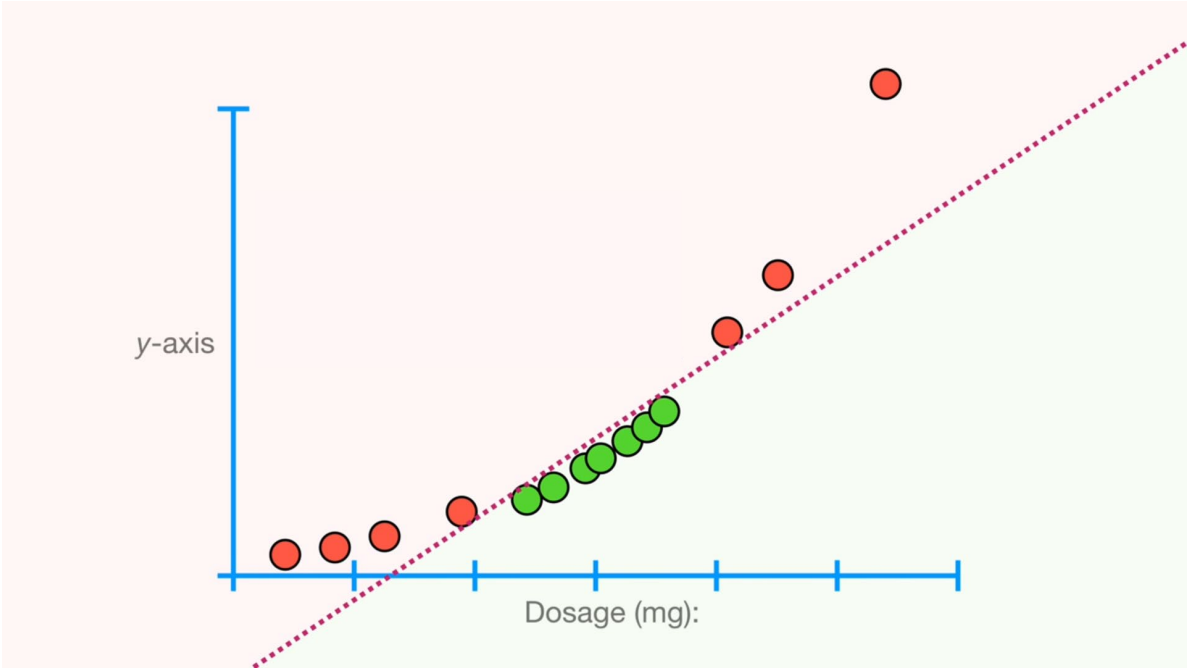
- Алгоритм был протестирован на нескольких типичных датасетах из HEP и показал устойчивость и хорошую классификационную способность в режиме поиска аномалий.
- Верх – выделение т-канального рождения топ-кварка из СМ фона.
- Низ – выделение нейтральных токов из СМ фона.



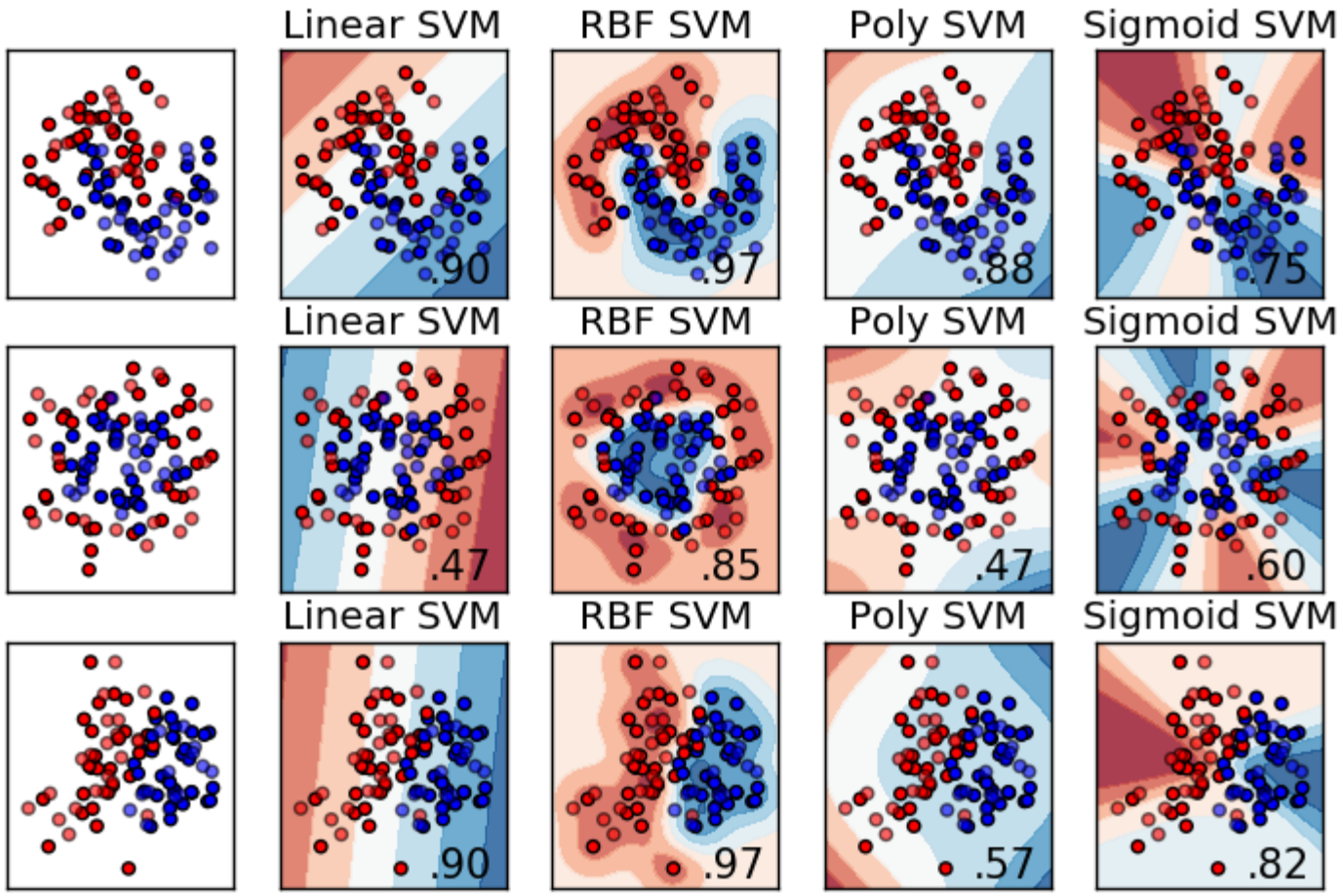



# Ядро Метода Опорных Векторов



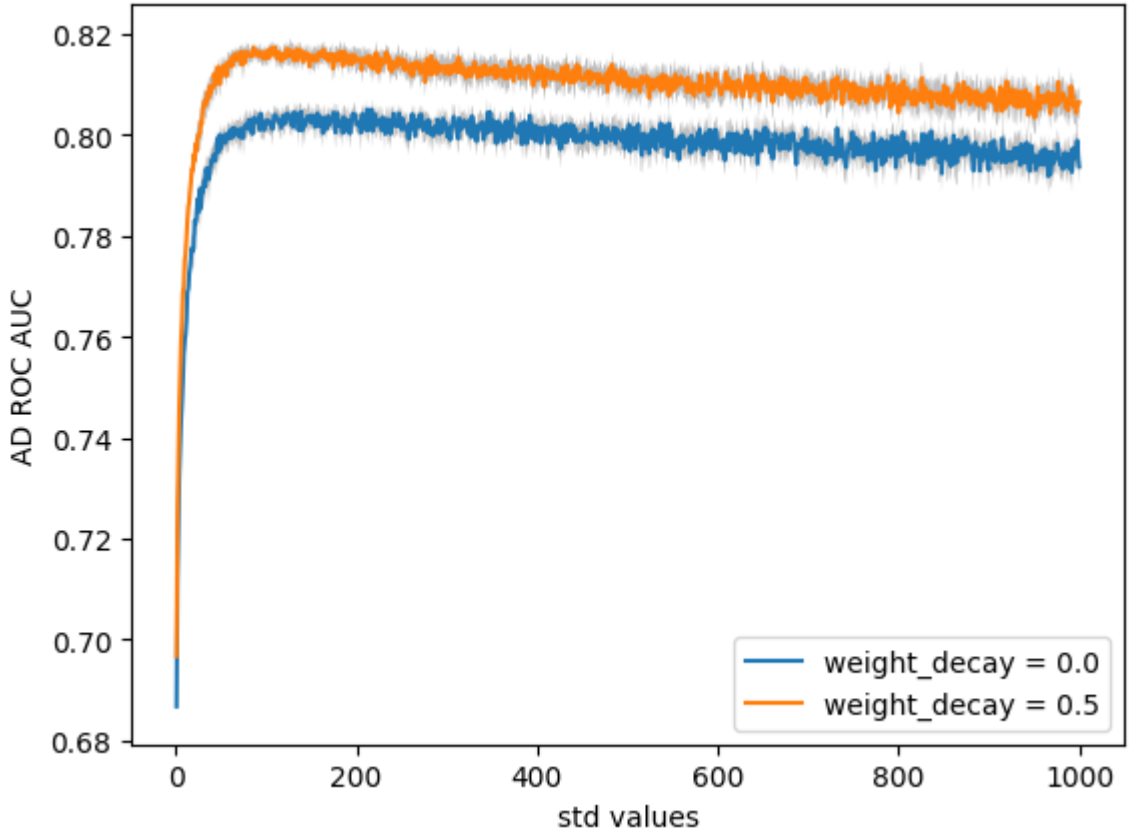



# Эффект ядер метода опорных векторов на поверхность решений



# Нейронная сеть для одного класса: параметры* шума

- Несмотря на то, что аномальный класс распределен вокруг нуля с std=1 (данные стандартизируются), оптимальные характеристики шума для алгоритма – это относительно большое стандартное отклонение.
- Это свидетельствует о том, что алгоритм действительно создает гиперповерхность вокруг нормального класса, а не шум удачным образом ложится на аномальный класс.

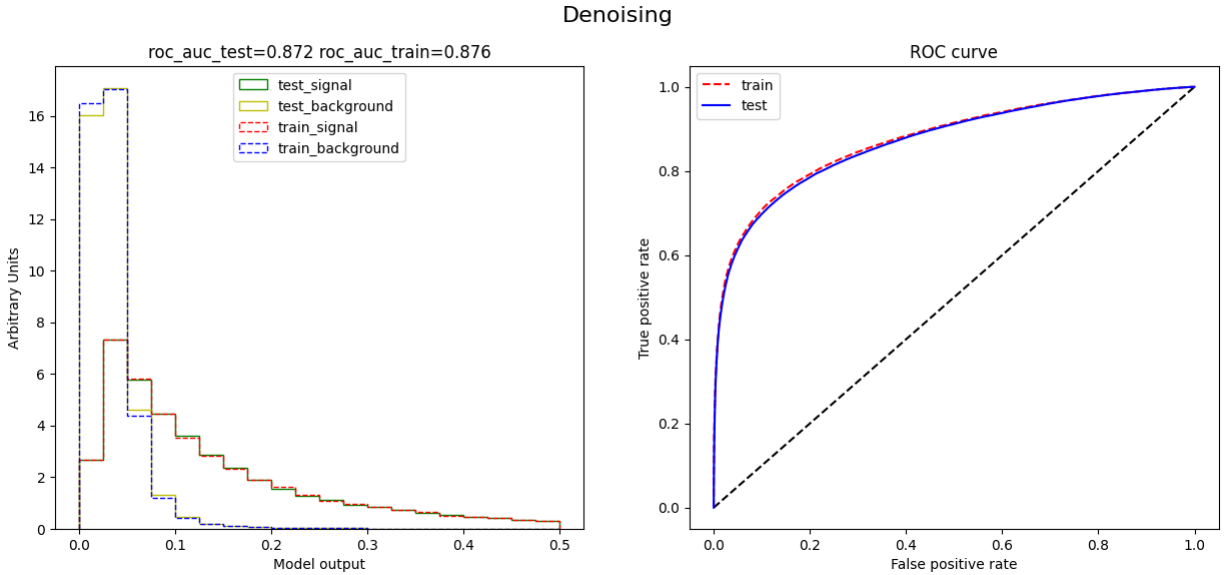



Зависимость точности алгоритма от стандартного отклонения для сгенерированного шума.

*Lev Dudko, P. V. Volkov, Georgi Vorotnikov и Andrei Zaborenko. «Application of Deep Learning Technique to an Analysis of Hard Scattering Processes at Colliders». В: Proceedings of The 5th International Workshop on Deep Learning in Computational Physics — PoS(DLCP2021) (2021).

# Метод детектирования аномалий с помощью удаления шума

- Метод основан на решении задачи удаления шума (денойзинга).
- К входным данным добавляется шум с небольшим стандартным отклонением (порядка 0.01 – 0.001). Модель реконструирует исходные переменные по зашумленным. Средняя ошибка реконструкции переменных принимается за метрику аномальности.




Модель работает на уровне KNN, лучше автоэнкодера, но хуже, чем OCDNN