Сравнительный анализ методов машинного и глубокого обучения в задаче классификации волновых форм полного электронного содержания

А.С. Тен[1], А.А. Сорокин[1], Н.В. Шестаков[2]

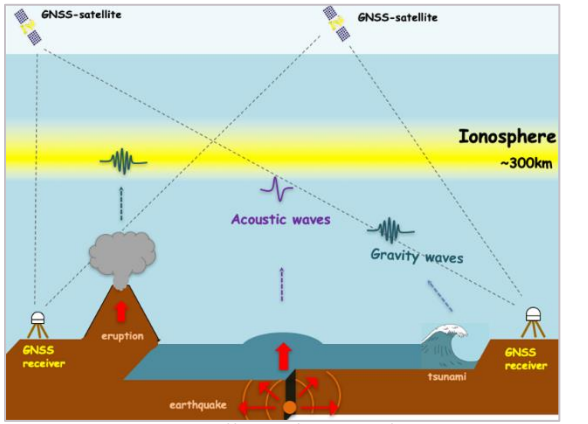
[1]ВЦ ДВО РАН, Хабаровск

[2]ДВФУ \ ИПМ ДВО РАН, Владивосток

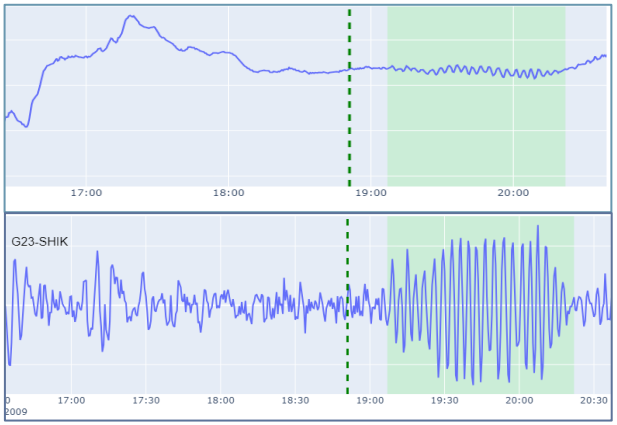
The 8th International Conference on Deep Learning in Computational Physics
June 19-21, 2024
SINP MSU, Moscow, Russia

Актуальность

- Извержения вулканов (VEI 2-6) порождают в ионосфере Земли ковулканические ионосферные возмущения (КИВ)[1].
- Для исследования КИВ анализируются временные ряды полного электронного содержания (ПЭС).
- Ряды ПЭС реконструируют из данных глобальных навигационных спутниковых систем (ГНСС) – GPS, ГЛОНАСС, «Бэйдоу», «Галилео».



источник изображения: https://eos.org/editors-vox/detecting-earths-natural-hazards-high-up-in-the-sky



[1] Astafyeva E. Ionospheric Detection of Natural Hazards // Reviews of Geophysics. 2019. № 4 (57). С. 1265–1288.

Актуальность

- Большой объем данных ГНСС.
- Существующие методы для автоматизированного поиска КИВ (STA\LTA, DROT[1]) не показывают высокой эффективности.
- Много операций с данными – высокий риск ошибок.

[1.] Karatay S. Detection of the ionospheric disturbances on GPS-TEC using differential rate of TEC (DROT) algorithm // Advances in Space Research. 2020. № 10 (65). С. 2372–2390.

Возможные решения

- Классификация временных рядов ПЭС.
- Методы машинного обучения и нейронные сети для анализа временных рядов уже успешно применяются в сейсмологии[1], метеорологии[2] и др. областях.
- Новые вычислители и адаптированный стек ПО.

1. Stepnov A., Chernykh V., Konovalov A. The seismo-performer: a novel machine learning approach for general and efficient seismic phase recognition from local earthquakes in real time // Sensors. 2021. № 18 (21).
2. Андреев А.И., Шамилова Ю.А. Детектирование облачности по данным КА Himawari-8 с применением сверточной нейронной сети // Исследование Земли из космоса. 2021. №2. с. 42-52.

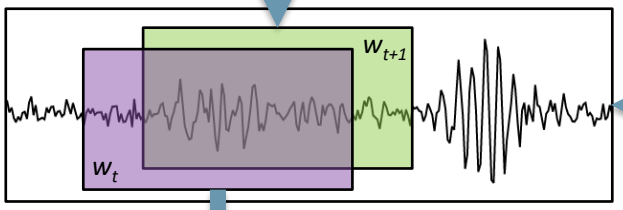
Цель исследования

Оценка применимости и сравнение некоторых методов классического машинного обучения и искусственных нейронных сетей в задаче бинарной классификации волновых форм ПЭС.

Основные задачи:

1. Сформировать обучающие наборы данных.
2. Сконструировать эффективные признаки данных исходных временных рядов для обучения классических моделей машинного обучения.
3. Обучить выбранные нейронные сети и классические модели машинного обучения, сравнить метрики качества классификации.
4. Сравнить обученные модели в автоматизированном поиске КИВ отложенных ГНСС-данных.

Алгоритм поиска ковулканических ионосферных возмущений



2.

Временные ряды ПЭС

1.

Архив ГНСС данных

Классификатор

3.

Классификатор:
- случайный лес
- град. бустинг (XGBoost)
- нейронная сеть

нет

Класс «КИВ» ?

да

4.

Сохранение времени начала и конца окна

Обнаруженные возмущения

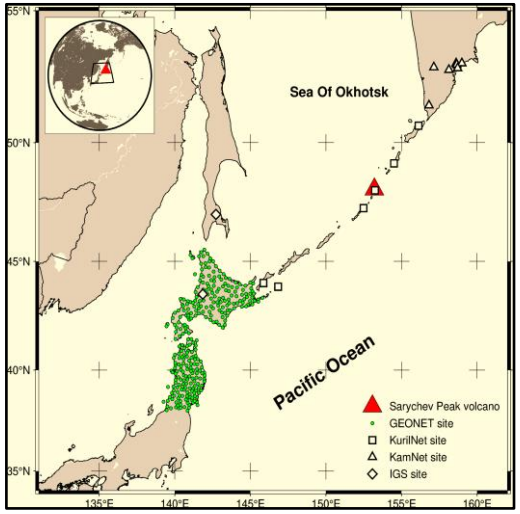
Классификаторы

Классификатор	Тип	Данные для обучения
MLP	Нейронная сеть	Исходные временные ряды ПЭС
FCN		
ResNet		
InceptionTime		
TSTransformer		
Случайный лес	Классич. модель машинного обучения	Признаки вычисленные из врем. рядов
Градиентный бустинг (XGBoost)		

1. Формирование обучающих наборов

источник: http://esgeo.ru/

источник: earthobservatory.nasa.gov



Временные ряды ПЭС с выделенными ковулканическими возмущениями
(пунктир. линия - эруптивное событие - 18:51 UT 14 июня 2009 г.)
а) исходные значения ПЭС; б) фильтрованные значения;

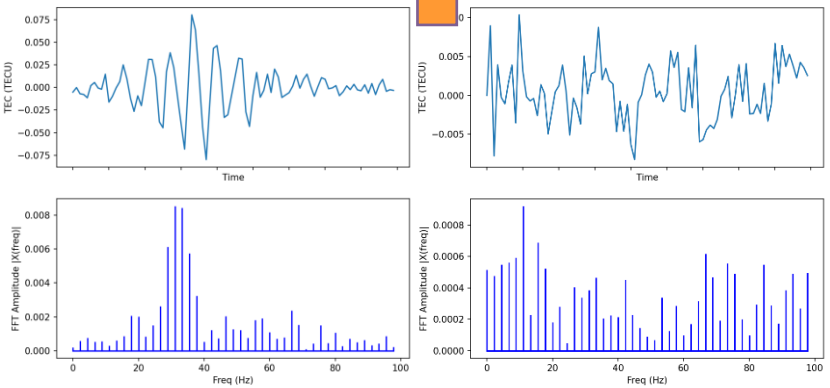
- Извержение вулкана Пик Сарычева (Большая Курильская гряда) 15 и 16 июня 2009 г., VEI = 4 [1] .
- Подготовлено 18 958 записей временных рядов ПЭС за 15 и 16 июня 2009 г. (японская сеть GEONET)
- Данные размечены – выделены КИВ [1].
- Сформирована обучающая выборка с образцами 2-х классов: «шум» (18441) и «КИВ» (318). (образец – в.р. 45 мин)
- 200 тестовых файлов отложено.

1. Shestakov N. [и др.]. Investigation of ionospheric response to June 2009 Sarychev peak volcano eruption // Remote Sensing. 2021. № 4 (13).

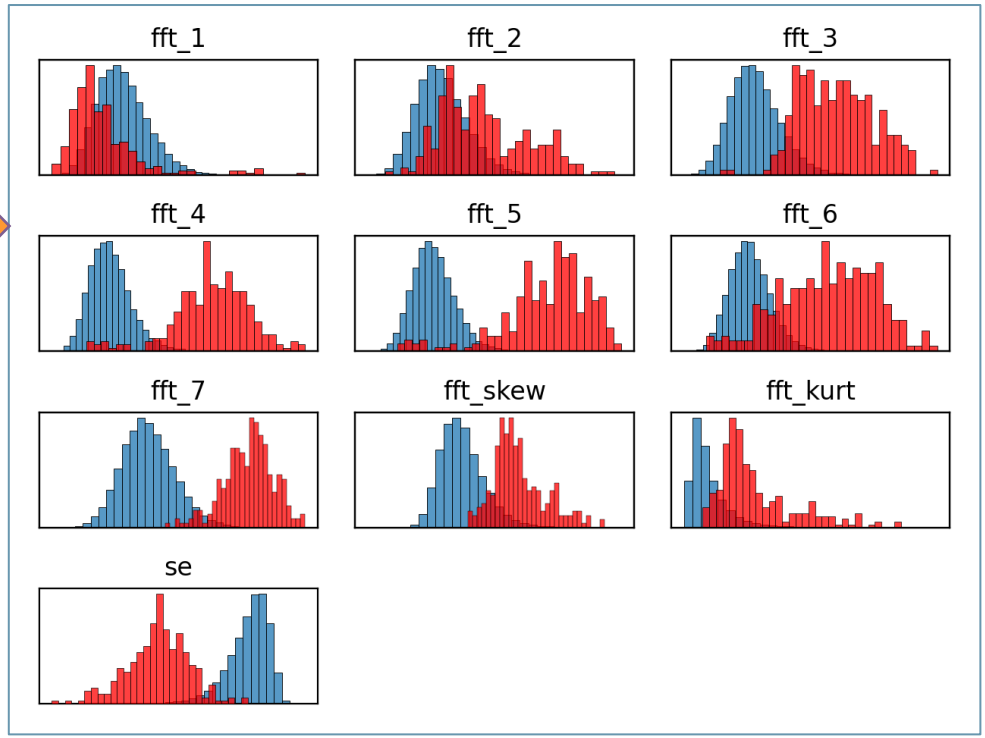
2. Конструирование признаков данных

Обозначение признака	Описание
fft_1 – fft_7	Вычислены по формуле (1)
se	Спектральная энтропия
fft_kurt	Коэффициент эксцесса
fft_skew	Коэффициент асимметрии

$$fft_{feature} = \frac{\sum_{f=f_{start}}^{f_{end}} X(f)}{\sum_{f=f_1}^{f_N} X(f)}, \quad (1)$$



Образцы класса «КИВ» и «шум»

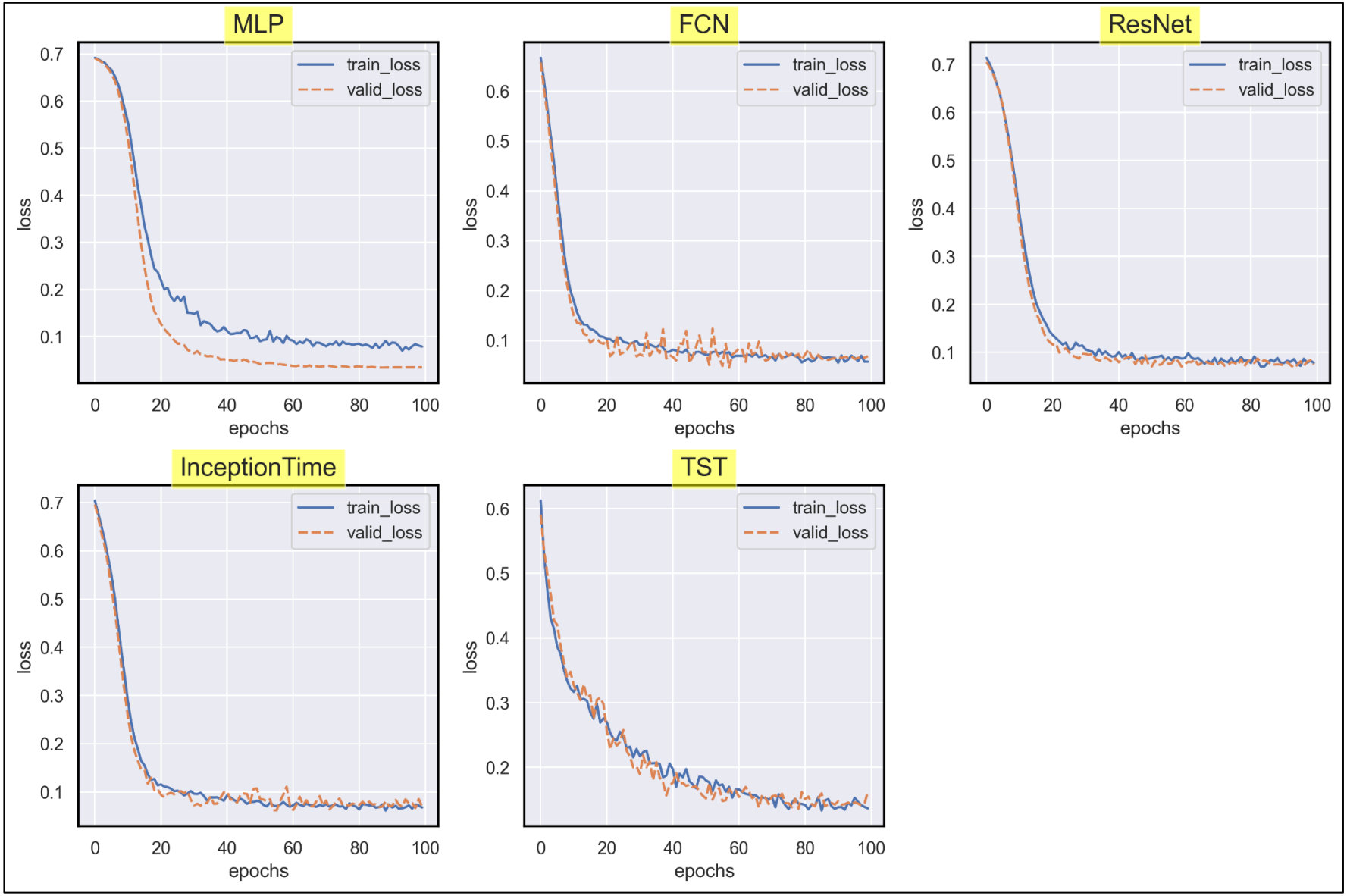


Гистограммы распределения значений предложенных признаков для каждого класса: синий – класс «шум», красный – класс «КИВ».

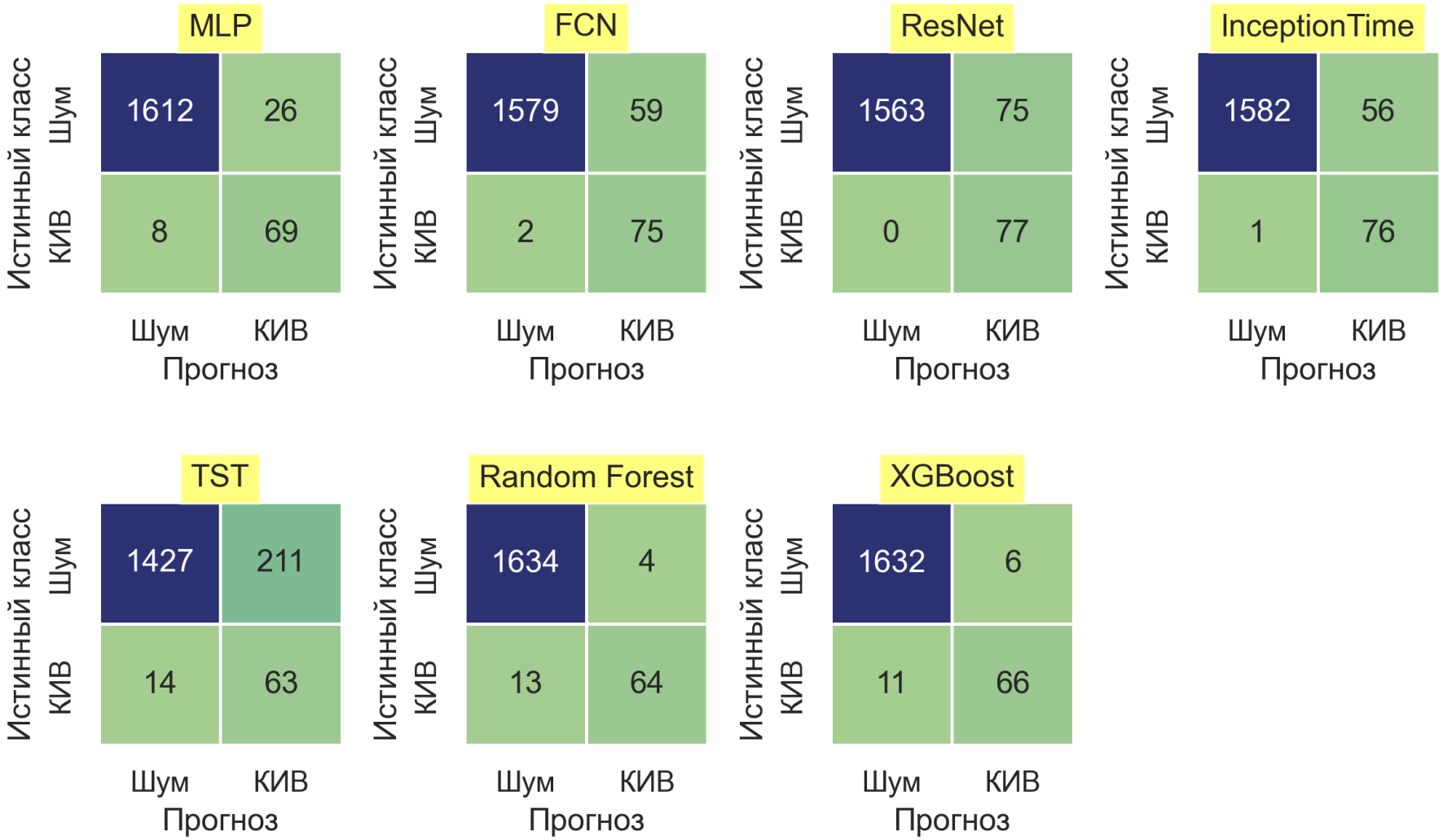
3. Обучение нейронных сетей

- Обучающий набор данный имеет сильный дисбаланс классов «шум» 18441, «КИВ» 318.
- Увеличение минорного класса исходной обучающей выборки до 25% выборки (upsampling) – SMOTE.
- Аугментация (гауссовский шум).
- Веса классов - шум: 0.5, КИВ: 30

3. Обучение нейронных сетей



3. Обучение классификаторов и оценка качества классификации



3. Обучение классификаторов и оценка качества классификации

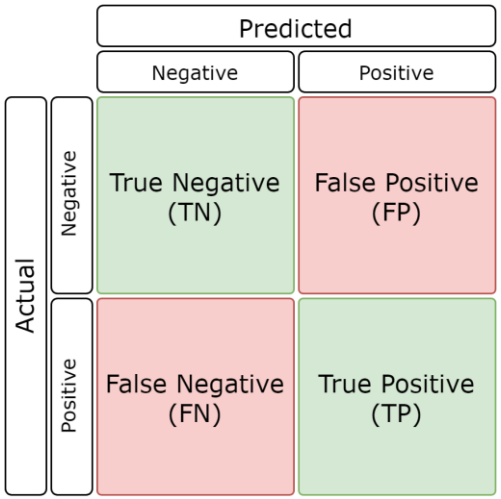
Модель \ метрика	Обучающая выборка		Тестовая выборка	
	F1	MCC	F1	MCC
MLP	0.96±0.01	0.95±0.01	0.90	0.80
FCN	0.93±0.01	0.91±0.01	0.85	0.72
ResNet	0.92±0.01	0.91±0.01	0.82	0.70
InceptionTime	0.93±0.01	0.91±0.0	0.85	0.74
TSTransformer	0.85±0.01	0.81±0.01	0.64	0.39
Случайный лес	0.89±0.02	0.79±0.04	0.94	0.88
Градиентный бустинг	0.87±0.01	0.75±0.01	0.95	0.90

Метрики качества классификации

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$F_1 = 2\frac{recall \cdot precision}{recall + precision} = \frac{TP}{TP + (FP + FN)/2}$$

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN}$$



4. Оценка эффективности классификаторов в алгоритме поиска КИВ

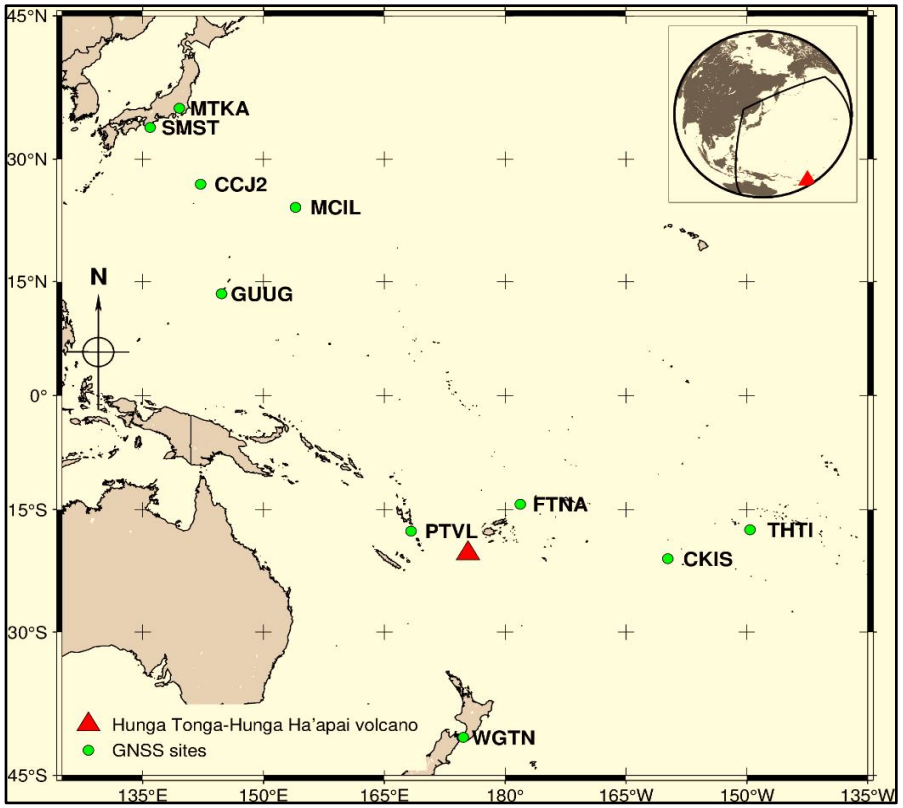
- **200** размеченных тестовых файлов ПЭС
- **73** из них – с КИВ (по 1 КИВ в файле)

Классификатор	Найдено КИВ (из 73)	Число ложных срабатываний (на 200 файлов)	Время работы алгоритма (сек) *
InceptionTime	70 (96%)	131 (0.66 в среднем на 1 файл)	28.5
FCN	70 (96%)	149 (0.75 в среднем на 1 файл)	25.4
MLP	68 (93%)	141 (0.71 в среднем на 1 файл)	24.1
Градиентный бустинг	**67 (92%)**	**36 (0.18 в среднем на 1 файл)**	**42.9**
STA/LTA	8 (11%)	90 (0.45 в среднем на 1 файл)	16.6

* x64-based PC, AMD Ryzen 7 4700U 8 cores 2.00 GHz, 16Gb RAM, 5 workers

Апробация классификаторов

Классификатор	Найдено КИВ (из 32)	Число ложных срабатываний (на 336 файлов)	Время работы алгоритма (сек) *
Градиентный бустинг	27	79 (0.25 в среднем на 1 файл)	52.7
InceptionTime	23	84 (0.25)	30
FCN	28	130 (0.39)	29.4
STA/LTA	13	846 (2.52)	20.1



- Поиск КИВ ассоциированных с извержением подводного вулкана Хунга-Тонга-Хунга-Хаапай (архипелаг Тонга), 15 января 2022 г. 4 ч. 02 мин по UTC, VEI 5-6[1]
- Классификатор обучен на данных по извержению влк. Пик Сарычева.
- 336 размеченных тестовых файлов ПЭС.
- 32 из них – с КИВ (по 1 КИВ в файле).

1. Yuen D. A. [и др.]. Under the surface: Pressure-induced planetary-scale waves, volcanic lightning, and gaseous clouds caused by the submarine eruption of Hunga Tonga-Hunga Ha'apai volcano // Earthquake Research Advances. 2022. № 3 (2). С. 100134.

* x64-based PC, AMD Ryzen 7 4700U 8 cores 2.00 GHz, 16Gb RAM, 5 workers

Заключение

- Показана применимость методов машинного и глубокого обучения в задаче поиска ковулканических ионосферных возмущений.
- Лучший результат – классификатор на основе градиентного бустинга, по соотношению числа найденных КИВ и ложных срабатываний в алгоритме поиска – 27 найденных КИВ из 32, ассоциированных с изв. влк. Хунга-Тонга-Хунга-Хаапай.
- Предложенные признаки данных показали эффективность для обучения исследуемых классификаторов, обеспечивают значение метрики MCC=0.90 на тестовой выборке (градиентный бустинг).

Спасибо за внимание!

Александр Тен
лаборатория информационных и
вычислительных систем ВЦ ДВО РАН
alexander.s.ten@yandex.ru