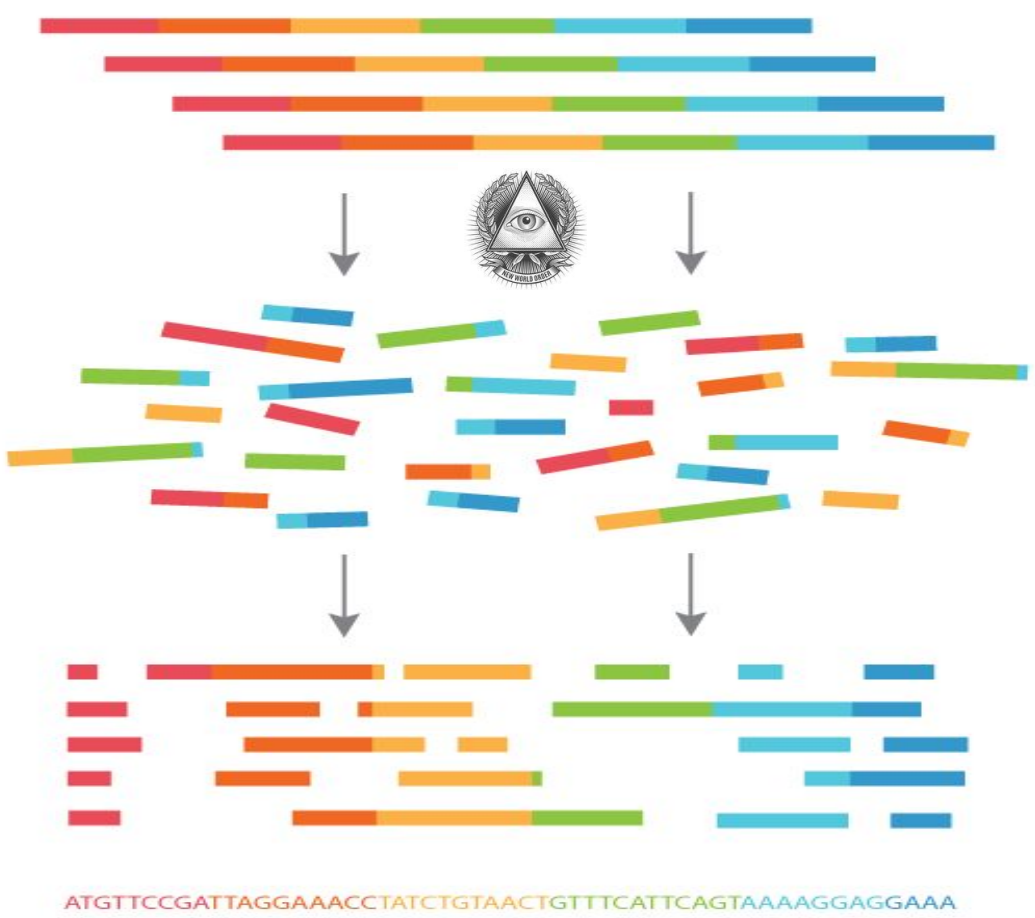
Подходы машинного обучения к определению новых бактериальных патогенов по данным NGS. Предсказание патогенности бактерии по данным NGS

Руководитель –
Константин Зайцев
Великий (Университет

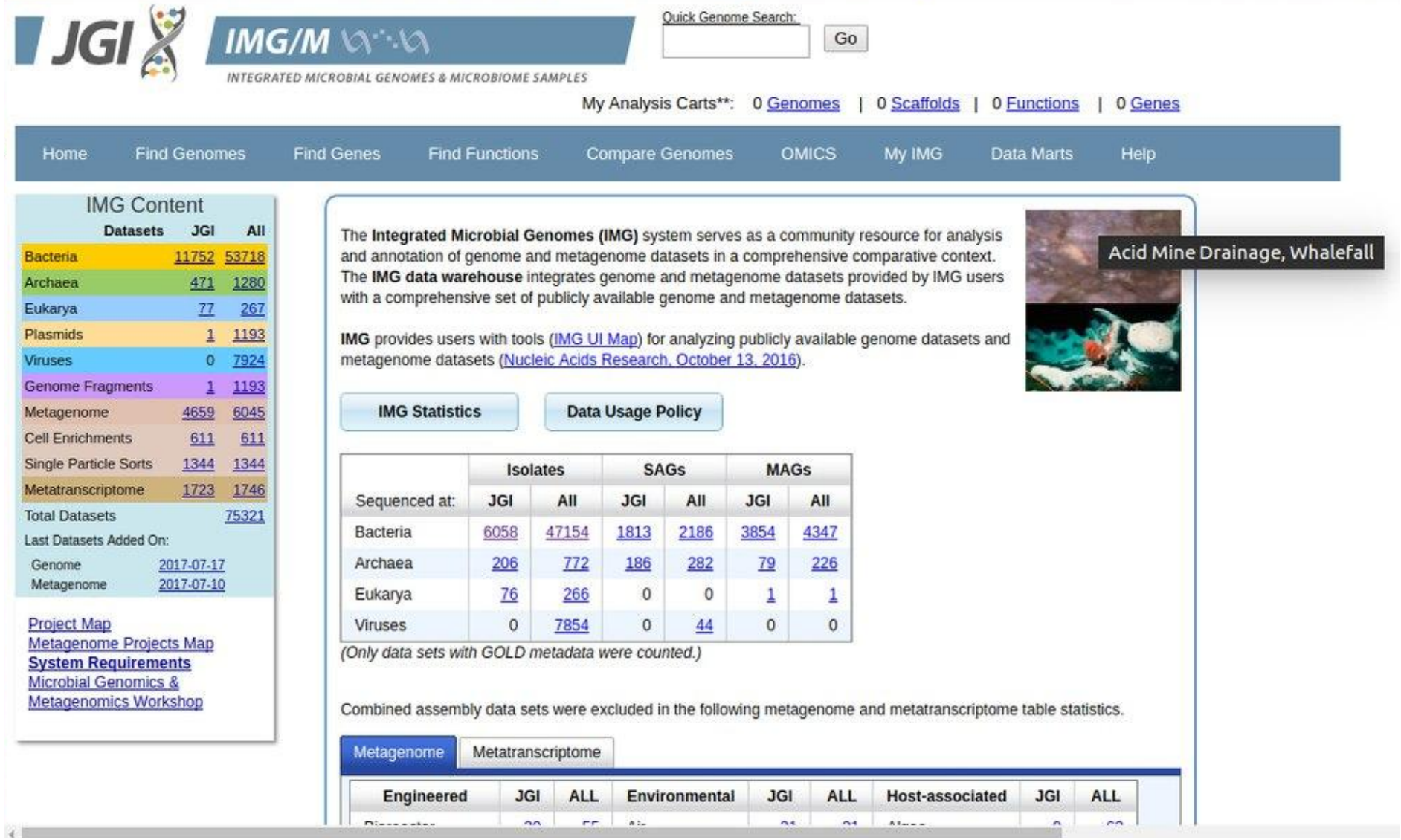
Цель работы

Использование подходов машинного обучения для создания предиктора патогенности бактерии по данным NGS-секвенирования, путем обучения на большом количестве штаммов с известной патогенностью и извлечением правильных признаков из данных NGS-секвенирования



Данные

- Для обучения использовались данные из таких баз как например

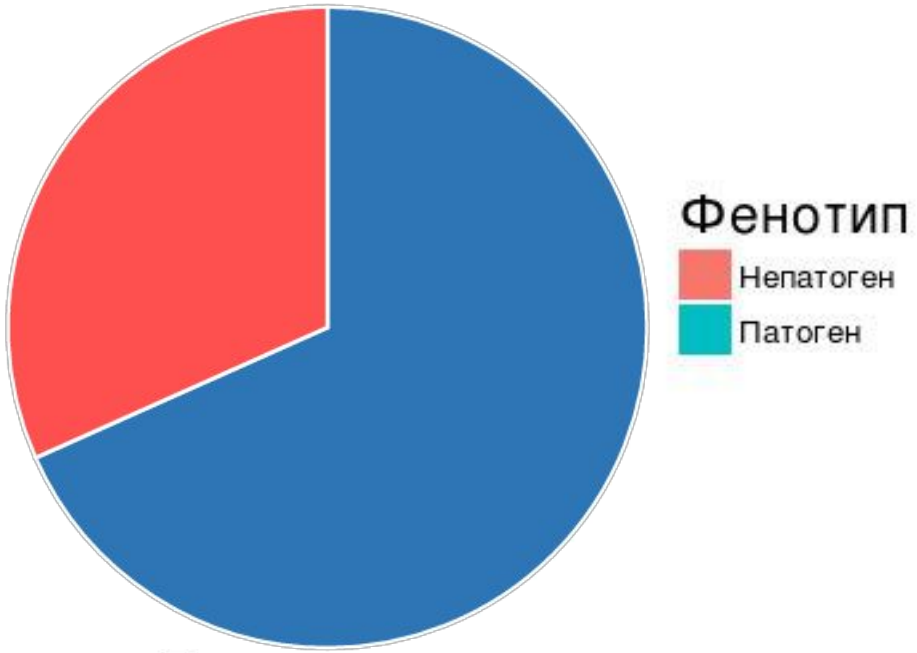


Классификация бактерий из выборки

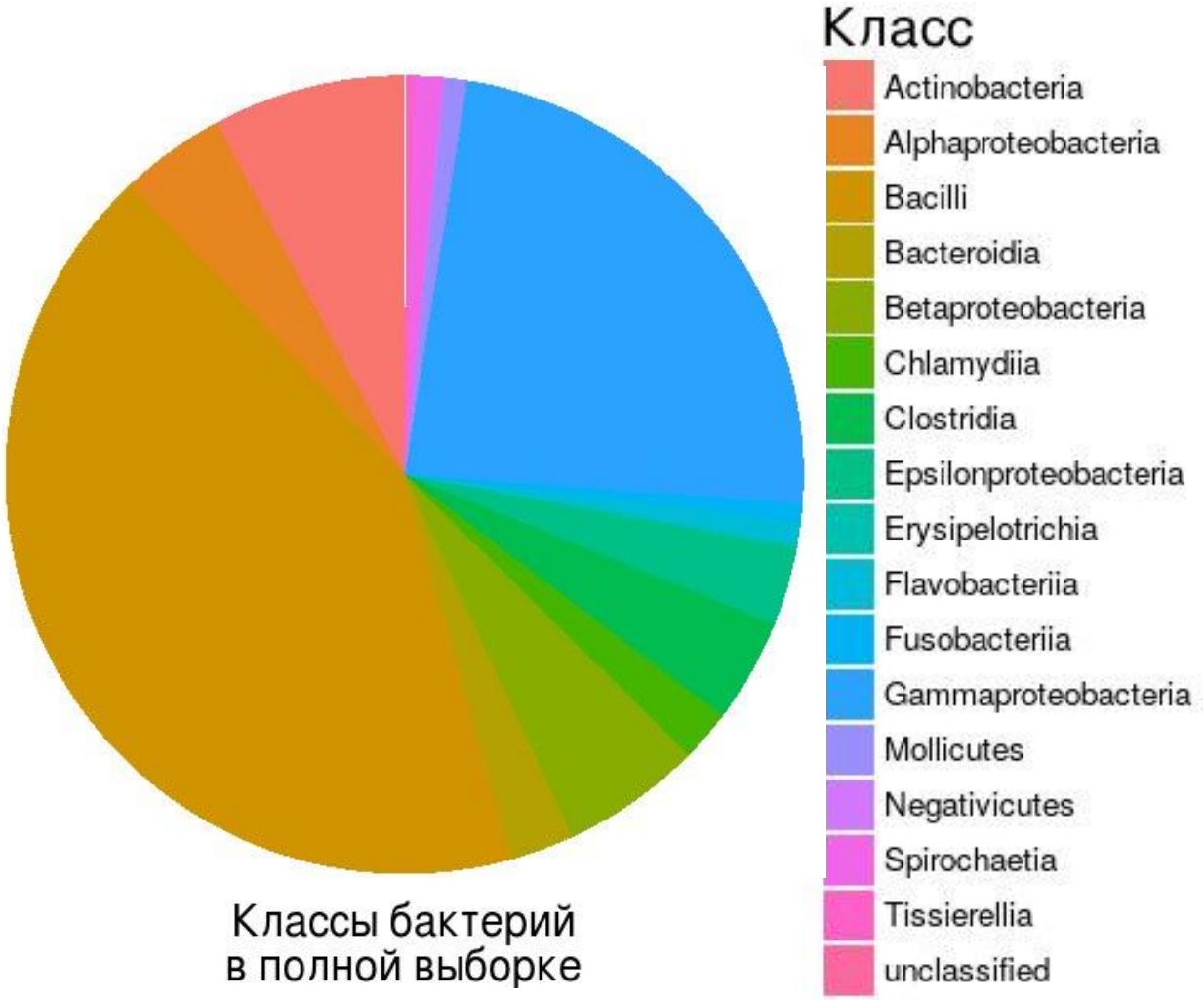
NP 418

P 902

1320



Доля патогенов
в выборке

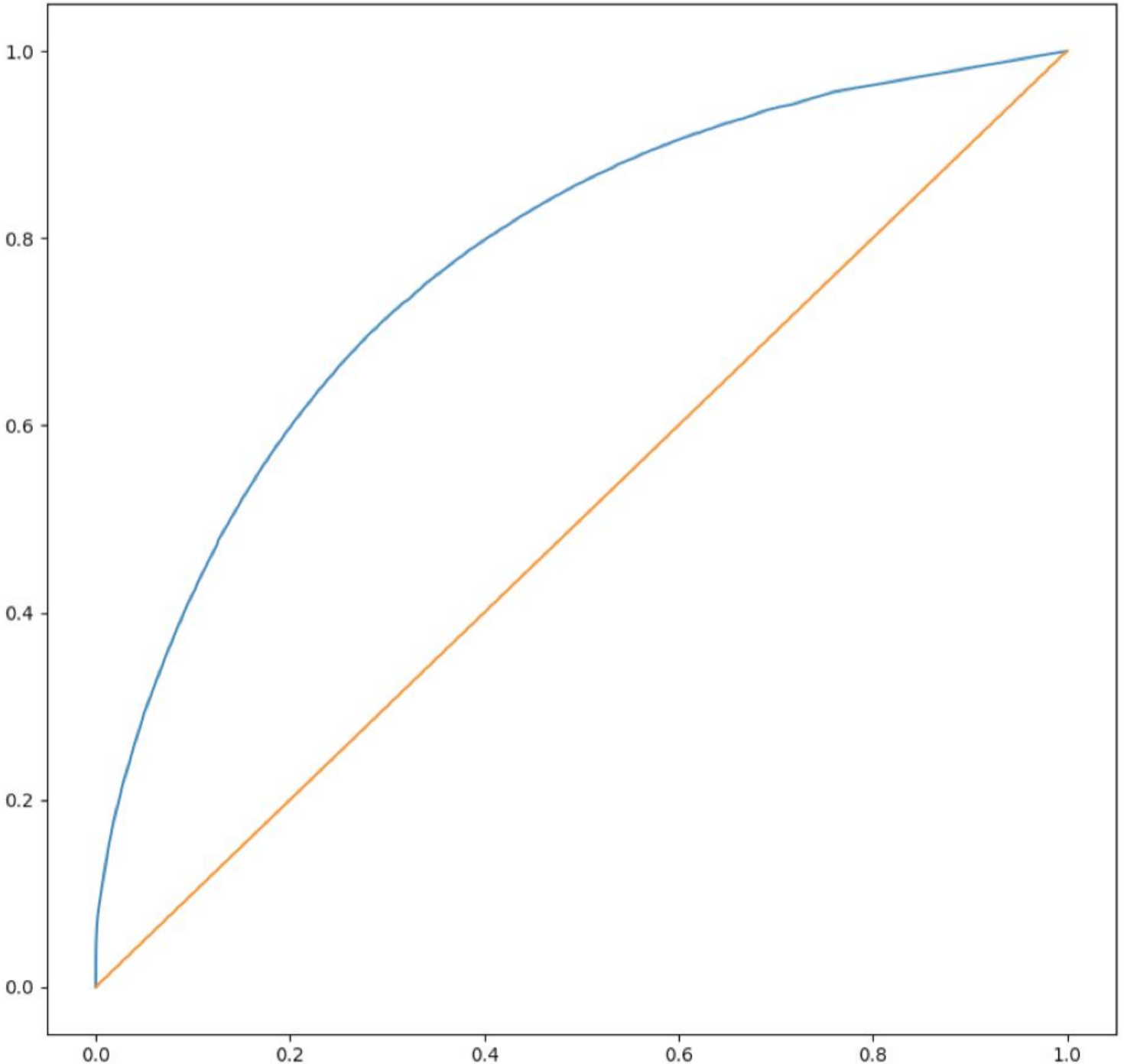


Классы бактерий
в полной выборке

Модели

Данные анализировали при помощи нескольких подходов –

1) Обучение сверточной нейронной сети для классификации отдельных ридов
2) Логистическая регрессия для классификации полных бактериальных геномов
3) Random Forest для классификации как отдельных ридов, так и полных бактериальных геномов



Результаты

AUC-ROC = 0.78

Логистическая регрессия для бактерий