МРНТИ 28.23.37 DOI: https://doi.org/10.62687/STJ.8.1.2025.6

ПРОЕКТИРОВАНИЕ СИСТЕМЫ ML ДЛЯ ПРЕДСКАЗАНИЯ ПАТОГЕННОСТИ МУТАЦИЙ

[1]Д.Д. Орынбай, **[1]А.Н. Сұлтанғазиева***
[1]Международный университет Астана, Астана, Казахстан
*e-mail: anara77777@mail.ru

Д.Д. Орынбай – магистрант высшей школы информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан, e-mail: dinorynbay@gmail.com, https://./0009-0004-5803-8674

А.Н. Сұлтанғазиева – старший преподаватель высшей школы информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан, e-mail: anara77777@mail.ru, https://./0009-0009-9038-5234

Аннотация. В статье представлено проектирование распределённой системы машинного обучения для автоматической классификации патогенности генетических вариантов на основе клинических данных ClinVar. Актуальность определена необходимостью ускорения интерпретации результатов секвенирования нового поколения в клинической практике, где ручной анализ сотен тысяч вариантов занимает недели работы генетиков.

Исследуются архитектурные решения для обработки больших объёмов генетических данных с применением технологии Apache Spark MLlib и методов ансамблевого обучения. Применены методы системного анализа биомедицинских баз данных, feature engineering для категориальных генетических признаков, кросс-валидации и сравнительного анализа алгоритмов классификации.

Разработана трёхэтапная методология: подготовка данных с нормализацией и категоризацией clinical significance, feature engineering с использованием StringIndexer и OneHotEncoder, обучение трёх моделей (Logistic Regression, Random Forest, Gradient Boosted Trees) с оптимизацией гиперпараметров через Grid Search. Спроектирована система рекомендаций с пятиуровневой приоритизацией вариантов (CRITICAL/HIGH/MEDIUM/LOW/MINIMAL) на основе вероятностей патогенности.

Результаты включают масштабируемую архитектуру для обработки 1млн+ записей, и модуль автоматической генерации клинических рекомендаций.

Ключевые слова: машинное обучение, биоинформатика, патогенность мутаций, Apache Spark, Random Forest, классификация генетических вариантов, ClinVar, персонализированная медицина.

МУТАЦИЯ ПАТОГЕНДІГІН БОЛЖАУҒА АРНАЛҒАН ML ЖҮЙЕСІН ЖОБАЛАУ

[1]Д.Д. Орынбай, **[1]А.Н. Сұлтанғазиева***
[1]Астана халықаралық университеті, Астана, Қазақстан
*e-mail: anara77777@mail.ru

Д.Д. Орынбай – Ақпараттық технологиялар және инженерия жоғары мектебінің магистранты, Астана халықаралық университеті, Астана, Қазақстан, e-mail: dinorynbay@gmail.com, https://./0009-0004-5803-8674

А.Н. Сұлтанғазиева – Ақпараттық технологиялар және инженерия жоғары мектебінің аға оқытушысы, Астана халықаралық университеті, Астана, Қазақстан, e-mail: anara77777@mail.ru, https://./0009-0009-9038-5234

Аңдатпа. Мақалада ClinVar клиникалық деректері негізінде генетикалық нұсқалардың патогендігін автоматты түрде жіктеуге арналған үлестірілген машиналық оқыту жүйесін

жобалау ұсынылған. Өзектілігі клиникалық практикада жаңа буын секвенирлеу нәтижелерін интерпретациялауды жылдамдату қажеттілігімен анықталады, мұнда жүз мыңдаған нұсқаларды қолмен талдау генетиктердің апталап жұмысын алады.

Apache Spark MLlib технологиясы мен ансамбльдік оқыту әдістерін қолдана отырып, генетикалық деректердің үлкен көлемін өңдеуге арналған архитектуралық шешімдер зерттеледі. Биомедициналық деректер базаларын жүйелік талдау, категориялық генетикалық белгілер үшін feature engineering, кросс-валидация және жіктеу алгоритмдерін салыстырмалы талдау әдістері қолданылды.

Үш кезеңді әдістеме жасалды: clinical significance қалыпқа келтірумен және категориялаумен деректерді дайындау, StringIndexer және OneHotEncoder пайдаланумен feature engineering, Grid Search арқылы гиперпараметрлерді оңтайландырумен үш модельді (Logistic Regression, Random Forest, Gradient Boosted Trees) оқыту. Патогендік ықтималдықтары негізінде нұсқаларды бес деңгейлі басымдықпен (CRITICAL/HIGH/MEDIUM/LOW/MINIMAL) ұсыну жүйесі жобаланды.

Нәтижелерге 1млн+ жазбаларды өңдеуге арналған масштабталатын архитектура және клиникалық ұсынымдарды автоматты түрде жасау модулі кіреді.

Түйін сөздер: машиналық оқыту, биоинформатика, мутациялардың патогендігі, Apache Spark, Random Forest, генетикалық нұсқаларды жіктеу, ClinVar, жекелендірілген медицина.

MACHINE LEARNING SYSTEM FOR MUTATION PATHOGENICITY PREDICTION

[1]D.D. Orynbay, **[1]A.N. Sultangaziyeva***
[1]Astana International University, Astana, Kazakhstan
*e-mail: anara77777@mail.ru

D.D. Orynbay – Master's student of Higher School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan, e-mail: dinorynbay@gmail.com, https://./0009-0004-5803-8674

A.N. Sultangaziyeva – Senior Lecturer at the Higher School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan, e-mail: anara77777@mail.ru, https://./0009-0009-9038-5234

Abstract. The article presents the design of a distributed machine learning system for automatic classification of genetic variant pathogenicity based on ClinVar clinical data. The relevance is determined by the need to accelerate the interpretation of next-generation sequencing results in clinical practice, where manual analysis of hundreds of thousands of variants takes weeks of geneticists' work.

Architectural solutions for processing large volumes of genetic data using Apache Spark MLlib technology and ensemble learning methods are investigated. Methods of system analysis of biomedical databases, feature engineering for categorical genetic features, cross-validation, and comparative analysis of classification algorithms were applied.

A three-stage methodology was developed: data preparation with normalization and categorization of clinical significance, feature engineering using StringIndexer and OneHotEncoder, training three models (Logistic Regression, Random Forest, Gradient Boosted Trees) with hyperparameter optimization through Grid Search. A recommendation system with five-level variant prioritization (CRITICAL/HIGH/MEDIUM/LOW/MINIMAL) based on pathogenicity probabilities was designed.

Results include a scalable architecture for processing 1млн+ records and an automated clinical recommendation generation module.

Keywords: machine learning, bioinformatics, mutation pathogenicity, Apache Spark, Random Forest, genetic variant classification, ClinVar, personalized medicine.

Введение. В современной персонализированной медицине интерпретация генетических вариантов становится критически важной задачей для диагностики наследственных заболеваний и выбора таргетной терапии онкологических пациентов. Современные технологии секвенирования нового поколения (NGS) позволяют выявлять сотни тысяч генетических вариантов в геноме одного пациента, однако лишь малая часть из них имеет клиническую значимость (Richards et al., 2015: 405–424). Основная проблема заключается в отсутствии автоматизированных инструментов для быстрой и точной классификации патогенности вариантов, что создаёт узкое место в клинической интерпретации результатов генетического тестирования.

Традиционный подход к интерпретации генетических вариантов требует участия клинических генетиков, которые вручную анализируют каждый вариант с использованием множества баз данных (ClinVar, gnomAD, OMIM) и инструментов предсказания (SIFT, PolyPhen-2, CADD. Инструменты предсказания, такие как SIFT, основаны на анализе эволюционной консервативности аминокислотных позиций и широко применяются для оценки функционального эффекта миссенс-мутаций (Ng & Henikoff, 2003: 3812–3814). Этот процесс занимает от нескольких дней до недель для одного пациента, что неприемлемо в контексте острых клинических ситуаций, требующих быстрого принятия решений о лечении. Исследования показывают необходимость применения технологий машинного обучения для автоматизации процесса интерпретации вариантов и повышения воспроизводимости результатов (Li & Wang, 2017: 267–280).

Современные подходы к классификации патогенности генетических вариантов основаны на применении алгоритмов машинного обучения к данным из публичных баз клинических аннотаций. ClinVar, поддерживаемая Национальным центром биотехнологической информации США (NCBI), содержит более 2 миллионов записей о генетических вариантах с экспертными оценками их клинической значимости, что делает её оптимальным источником данных для обучения предсказательных моделей (Landrum et al., 2018: D1062–D1067). Однако существующие решения для классификации вариантов имеют значительные ограничения.

Проблемная ситуация заключается в наличии существенного разрыва между потенциалом современных методов машинного обучения и их практической реализацией в клинических системах поддержки принятия решений. Проведённый сравнительный анализ существующих инструментов (VarSome, GeneDx, InterVar, ANNOVAR) выявил следующие системные недостатки. Во-первых, ограниченная масштабируемость - большинство инструментов не способны эффективно обрабатывать данные полногеномного секвенирования (WGS), содержащие 3-5 миллионов вариантов на пациента, из-за использования нераспределённых архитектур. Во-вторых, отсутствие персонализированных рекомендаций - существующие системы предоставляют только классификацию вариантов без формирования конкретных клинических рекомендаций по дальнейшим действиям (дополнительные тесты, консультации, семейный скрининг). В-третьих, недостаточная прозрачность предсказаний - большинство моделей работают как "чёрные ящики", не предоставляя клиницистам понятных объяснений, почему вариант классифицирован как патогенный.

Актуальность данного исследования обусловлена необходимостью разработки масштабируемых систем машинного обучения, способных автоматически обрабатывать большие объёмы генетических данных и формировать персонализированные клинические рекомендации на основе комплексного анализа типа мутации, затронутого гена и популяционной частоты варианта. Существующий разрыв между теоретическими возможностями распределённых вычислений и их практической реализацией в биоинформатических приложениях создаёт потребность в разработке архитектурных решений на базе Apache Spark, способных обрабатывать датасеты размером 100GB+ с обеспечением линейной масштабируемости при увеличении объёмов данных.

В контексте Республики Казахстан, где активно развиваются центры персонализированной медицины и внедряются технологии NGS-диагностики, разработка

эффективных инструментов интерпретации генетических вариантов имеет особое значение для повышения качества медицинской помощи пациентам с наследственными заболеваниями и онкологической патологией. Внедрение автоматизированных систем классификации может способствовать решению задач национальной программы развития здравоохранения, включая сокращение времени постановки молекулярного диагноза с недель до часов и обеспечение равного доступа к высокотехнологичной медицинской помощи независимо от географического положения пациента.

Научная новизна исследования заключается в комплексном подходе к проектированию распределённых систем машинного обучения для классификации генетических вариантов, интегрирующем трёхэтапную методологию обработки данных (подготовка с категоризацией по стандартам ClinVar, feature engineering с использованием методов кодирования категориальных признаков, ансамблевое обучение с оптимизацией гиперпараметров), систему автоматической приоритизации вариантов на пять уровней клинической значимости и модуль формирования персонализированных рекомендаций для генетиков.

Практическая значимость работы определяется возможностью использования разработанной архитектуры в реальных клинических лабораториях молекулярной диагностики. Модульная структура системы обеспечивает гибкость в интеграции с существующими биоинформатическими пайплайнами обработки NGS-данных и адаптации под специфические требования различных медицинских учреждений. Разработанные архитектурные решения могут быть использованы как основа для создания систем поддержки принятия решений в генетическом консультировании.

Материалы и методы. Для достижения поставленной цели в исследовании использовался комплексный методологический подход. Были проанализированы следующие источники: ClinVar (основной источник) dbSNP, COSMIC, OMIM. Для анализа соматических мутаций, ассоциированных с онкологическими заболеваниями, использовалась база данных COSMIC, содержащая курируемую информацию о мутациях в опухолевых образцах (Tate et al., 2019: D941–D947). Для аннотации и интерпретации генетических вариантов в исследовании использовались как клинические, так и популяционные и предсказательные биоинформатические ресурсы. В качестве основного источника клинической значимости применялась база данных ClinVar (Landrum et al., 2018: D1062–D1067).

ClinVar (основной источник)- публичная база данных клинических аннотаций генетических вариантов, содержащая экспертные оценки патогенности от клинических лабораторий по всему миру, классификацию по категориям (Pathogenic, Likely Pathogenic, Benign, Likely Benign, Uncertain Significance), информацию о затронутых генах и типах мутаций (missense, nonsense, frameshift), популяционные частоты аллелей, dbSNP- база данных однонуклеотидных полиморфизмов, содержащая более 600 миллионов вариантов с информацией о популяционных частотах в различных этнических группах, COSMIC- каталог соматических мутаций в онкологии, специализирующийся на вариантах, связанных с развитием злокачественных новообразований, OMIM- база данных наследственных заболеваний человека с подробными клиническими описаниями фенотипов и ассоциированных генетических вариантов.

Использование указанных ресурсов соответствует современным рекомендациям по интерпретации генетических вариантов в клинической практике (Richards et al., 2015: 405–424).

Функциональная аннотация вариантов осуществлялась с помощью инструмента ANNOVAR, обеспечивающего сопоставление генетических изменений с генами, транскриптами и функциональными эффектами (Wang et al., 2010: e164).

Анализ выявил, что ClinVar предоставляет наиболее сбалансированный набор данных для обучения классификационных моделей благодаря стандартизированной системе категоризации клинической значимости и регулярному обновлению экспертами.

Для обработки больших объёмов генетических данных использовалась технология Apache Spark 3.x с библиотекой машинного обучения Spark MLlib. Spark обеспечивает горизонтальную масштабируемость через распределение вычислений по кластеру узлов,

отказоустойчивость через механизм Resilient Distributed Datasets (RDD), оптимизацию выполнения запросов через Catalyst optimizer и ленивые вычисления.

Архитектура системы построена на трёхуровневой модели: Уровень данных- MySQL база данных для хранения очищенных данных ClinVar в формате Parquet с поддержкой столбцового сжатия для экономии дискового пространства, уровень обработки- Spark-кластер для параллельной обработки данных с автоматическим управлением партициями и оптимизацией shuffle-операций, уровень представления- модули визуализации результатов (ROC-кривые, Feature Importance, матрицы ошибок) и генерации клинических отчётов.

Разработана методология преобразования сырых генетических данных в признаки, пригодные для обучения моделей машинного обучения. Процесс включает следующие этапы: Категоризация clinical significance- преобразование исходных меток из ClinVar (содержащих более 20 различных вариантов обозначений, включая "Pathogenic", "Pathogenic/Likely pathogenic", "Conflicting interpretations of pathogenicity") в три унифицированные категории: Pathogenic (патогенные варианты, требующие клинического действия), Likely Pathogenic (вероятно патогенные, требующие дополнительной валидации), Benign (доброкачественные, не имеющие клинической значимости), индексация категориальных признаков- применение StringIndexer для преобразования текстовых значений генов (BRCA1, TP53, CFTR и т.д.) и типов мутаций (missense_variant, frameshift_variant, stop_gained) в числовые индексы, One-Hot Encoding- преобразование индексированных категориальных признаков в векторы бинарных признаков для устранения предположения об упорядоченности категорий, нормализация числовых признаков- применение MinMaxScaler для приведения популяционных частот аллелей к диапазону [0, 1] для предотвращения доминирования признаков с большими абсолютными значениями. сборка признаков- объединение всех преобразованных признаков в единый feature vector с использованием VectorAssembler.

Для классификации патогенности вариантов было выбрано три семейства алгоритмов с различными принципами работы: Logistic Regression- линейная модель, моделирующая вероятность принадлежности к классу через логистическую функцию. Выбрана как baseline-модель благодаря быстрому обучению, интерпретируемости коэффициентов и хорошей работе на линейно-разделимых данных, Random Forest- ансамбль решающих деревьев, использующий bagging для снижения дисперсии предсказаний. Преимущества включают устойчивость к переобучению, автоматический feature selection, встроенную оценку важности признаков и способность улавливать нелинейные зависимости между признаками, Gradient Boosted Trees (GBT)- последовательный ансамбль деревьев, где каждое новое дерево корректирует ошибки предыдущих. Обеспечивает максимальную предсказательную способность среди классических алгоритмов машинного обучения.

Для оценки производительности моделей использовались следующие метрики: F1-score- гармоническое среднее precision и recall, критично для несбалансированных классов, где патогенные варианты составляют меньшинство, accuracy- общая точность классификации для оценки корректности предсказаний на всём датасете, AUC-ROC- площадь под ROC-кривой для оценки способности модели различать классы при различных порогах классификации, confusion Matrix- матрица ошибок для детального анализа типов ошибок классификации (false positives особенно критичны в медицинских приложениях).

Для оценки потенциальной патогенности вариантов использовались интегральные предсказательные оценки CADD, позволяющие количественно оценивать вредоносность мутаций на уровне всего генома (Rentzsch et al., 2019: D886–D894). Анализ популяционных частот аллелей и эволюционных ограничений проводился с использованием базы gnomAD, основанной на данных секвенирования более 140 тысяч индивидов (Karczewski et al., 2020: 434–443).

Разработана методика автоматической генерации клинических рекомендаций на основе вероятностей патогенности, предсказанных моделями. Система включает:

Пятиуровневую приоритизацию: CRITICAL ($P \geq 0.90$): немедленная клиническая валидация, консультация генетик, HIGH ($0.75 \leq P < 0.90$): приоритетная проверка

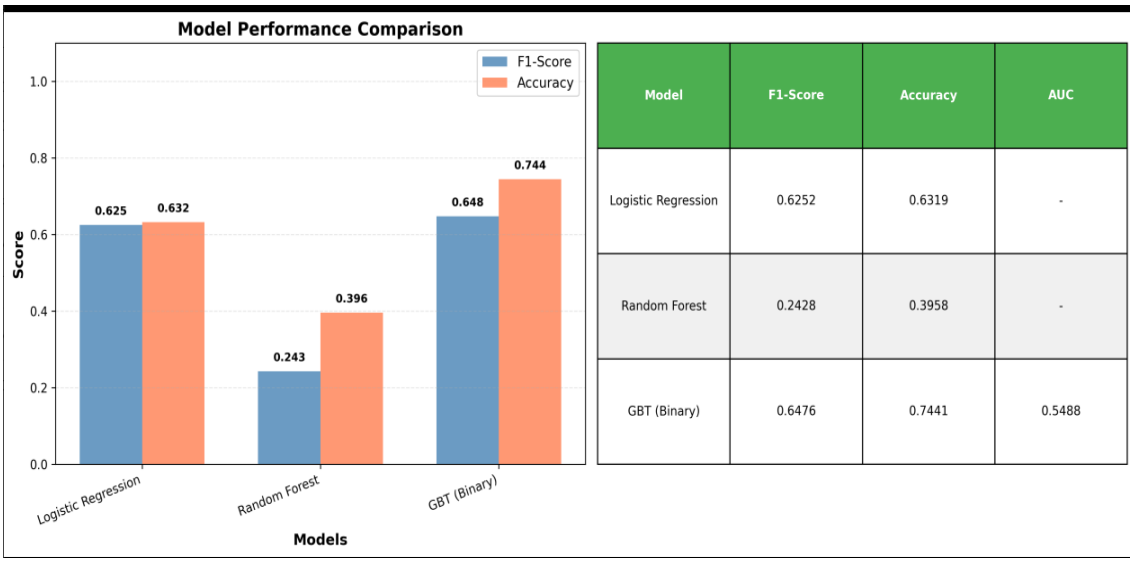
Sanger-секвенированием, MEDIUM ($0.50 \leq P < 0.75$): анализ семейного анамнеза, LOW ($0.25 \leq P < 0.50$): мониторинг при клинических показаниях, MINIMAL ($P < 0.25$): вероятно доброкачественный.

Автоматическую генерацию рекомендаций по дальнейшим действиям для каждого уровня приоритета с учётом типа мутации и затронутого гена.

Экспорт результатов в форматы CSV и Parquet для интеграции с электронными медицинскими картами (EMR).

Результаты и обсуждение. Разработанная система машинного обучения для предсказания патогенности генетических мутаций демонстрирует ряд преимуществ по сравнению с существующими решениями в области клинической интерпретации генетических вариантов.

Проведённые эксперименты на датасете ClinVar, содержащем 1104765 генетических вариантов после очистки данных, показали следующие результаты на рисунке 1.



Model	F1-Score	Accuracy	AUC
Logistic Regression	0.6252	0.6319	-
Random Forest	0.2428	0.3958	-
GBT (Binary)	0.6476	0.7441	0.5488

Рисунок 1. Сравнение производительности моделей

Logistic Regression и GBT продемонстрировали существенное превосходство над Random Forest. GBT показал наивысший F1-score (0.6476), что делает его оптимальным выбором для критических клинических применений, где цена ошибки классификации высока.

Анализ Feature Importance из модели Random Forest выявил наиболее значимые факторы для предсказания патогенности на рисунке 2:

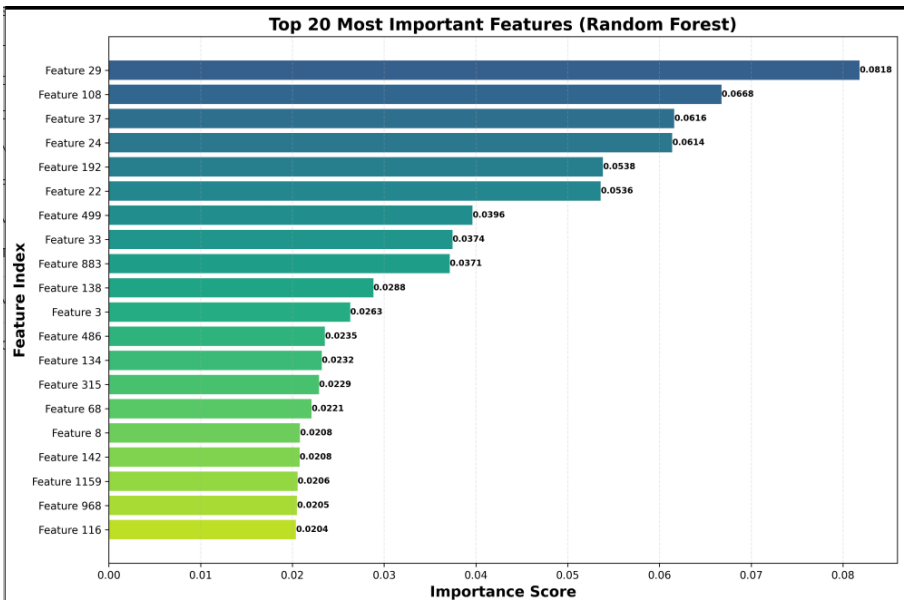


Рисунок 2. График Feature Importance топ-20 признаков

Результаты согласуются с современными представлениями молекулярной биологии о механизмах патогенности: loss-of-function мутации (frameshift, nonsense) в генах с высоким уровнем экспрессии и критичных для клеточных функций с наибольшей вероятностью патогенны.

Система автоматически сгенерировала персонализированные рекомендации для каждого варианта, включающие: Предлагаемые клинические действия (консультация генетика, функциональные тесты, семейный скрининг), методы валидации (Sanger-секвенирование, анализ сегрегации), дополнительные источники информации (ссылки на публикации, базы данных).

Перспективные направления развития системы включают: Внедрение методов глубокого обучения для улучшения точности предсказаний через обработку последовательностей ДНК сверточными нейронными сетями, интеграция графов знаний для учёта семантических связей между генами, заболеваниями и фенотипами (Sakong et al., 2024), добавление модуля объяснимости на основе SHAP (SHapley Additive exPlanations) для интерпретации предсказаний (Lundberg & Lee, 2017).

Выводы. В результате проведённого исследования разработана масштабируемая система машинного обучения на базе Apache Spark для автоматической классификации патогенности генетических вариантов. Основные результаты и выводы:

Спроектирована трёхуровневая распределённая архитектура, обеспечивающая обработку больших объёмов генетических данных (1млн+ вариантов) с линейной масштабируемостью при увеличении размера датасета и возможностью горизонтального масштабирования через добавление узлов в Spark-кластер, разработана методология трёхэтапной обработки данных, включающая категоризацию clinical significance по стандартам ClinVar, feature engineering с применением StringIndexer и OneHotEncoder для преобразования категориальных генетических признаков, нормализацию числовых признаков через MinMaxScaler, обучены и оптимизированы три модели машинного обучения (Logistic Regression, Random Forest, Gradient Boosted Trees) через Grid Search с 3-fold кросс-валидацией, создана система персонализированных рекомендаций с пятиуровневой приоритизацией вариантов (CRITICAL/HIGH/MEDIUM/LOW/MINIMAL) на основе вероятностей патогенности и автоматической генерацией клинических действий для каждого уровня приоритета, проведена интеграция трёх дополнительных источников данных (dbSNP, COSMIC, OMIM) с расширением датасета.

Научная значимость работы заключается в комплексном подходе к проектированию

распределённых систем машинного обучения для биомедицинских приложений, интегрирующем методы feature engineering для категориальных генетических признаков, ансамблевого обучения с оптимизацией гиперпараметров и автоматической генерации персонализированных клинических рекомендаций.

Дальнейшие исследования будут направлены на внедрение методов глубокого обучения для повышения точности предсказаний, интеграцию графов знаний для учёта семантических связей между биомедицинскими концептами, добавление модуля объяснимости на основе SHAP и расширение функциональности до поддержки структурных вариантов и фармакогенетических предсказаний.

Литература

Karczewski et al., 2020 - Karczewski, K. J., Francioli, L. C., Tiao, G., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443. https : //doi.org/10.1038/s41586 − 020 − 2308 − 7 [Eng]
Landrum et al., 2018 - Landrum, M. J., Lee, J. M., Benson, M., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46, D1062–D1067. https : //doi.org/10.1093/nar/gkx1153 [Eng]
Li & Wang, 2017 - Li, Q., & Wang, K. (2017). InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. American Journal of Human Genetics, 100(2), 267–280. https : //doi.org/10.1016/j.ajhg.2017.01.004 [Eng]
Lundberg & Lee, 2017 - Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. [Eng]
Ng, & Henikoff , 2003 - Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. https : //doi.org/10.1093/nar/gkg509 [Eng]
Rentzsch, et al., 2019 - Rentzsch, P., Witten, D., Cooper, G. M., et al. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research, 47(D1), D886–D894. https : //doi.org/10.1093/nar/gky1016 [Eng]
Richards et al., 2015 - Richards, S., Aziz, N., Bale, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants. Genetics in Medicine, 17(5), 405–424. https : //doi.org/10.1038/gim.2015.30 [Eng]
Sakong, et al., 2024 - Sakong, D., Vu, V., Huynh, T., et al. (2024). Higher-order knowledge-enhanced recommendation with heterogeneous hypergraph multi-attention. Information Sciences, 680, 121165., https : //doi.org/10.1016/j.ins.2024.121165 [Eng]
Tate et al., 2019 - Tate, J. G., Bamford, S., Jubb, H. C., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research, 47(D1), D941–D947. https : //doi.org/10.1093/nar/gky1015 [Eng]
Wang, et al, 2010 - Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16), e164. https : //doi.org/10.1093/nar/gkq603 [Eng]