УДК 004.8:61

АНАЛИЗ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Стукалова Д. А., студ., Мещеряк В. В., студ., Архипов И. Д., студ., Дунина Е. Б., к.ф.-м.н., доц.
Витебский государственный технологический университет, г. Витебск, Республика Беларусь

Реферат. *Актуальность исследования обусловлена возрастающей сложностью анализа генетических данных, особенно в контексте наследственных заболеваний, таких как ретинобластома, связанная с мутациями гена RB1. Традиционные методы интерпретации мутаций требуют значительных временных и экспертных ресурсов, что ограничивает их применение в клинической практике. Целью работы является разработка автоматизированного подхода на основе искусственного интеллекта для классификации патогенных вариантов RB1.*

В исследовании предложена архитектура свёрточной нейронной сети (CNN), адаптированная для обработки последовательностей ДНК. Модель обучена на данных из базы ClinVar, включающих аннотированные патогенные мутации. Для балансировки классов использован метод синтетической выборки. Обучение проводилось с метриками Accuracy, Precision, Recall и AUC, а также применением ранней остановки и динамического изменения скорости обучения для предотвращения переобучения.

Ключевые слова: искусственный интеллект, ген RB1, нейронные сети, генетический анализ, биоинформатика.

Современная геномика сталкивается с проблемой обработки экспоненциально растущих объемов данных. Особую сложность представляют мутации гена RB1, связанные с ретинобластомой – злокачественным заболеванием сетчатки, диагностируемым у детей до 5 лет [1, 2]. Традиционные методы анализа требуют значительных временных затрат и подвержены субъективным ошибкам.

Как показали Knudson (1971) и Goodrich (2006), инактивация RB1 приводит к нарушению регуляции клеточного цикла. Существующие биоинформатические инструменты (ANNOVAR, VEP) и методы машинного обучения (DeepSEA, CADD) позволяют прогнозировать патогенность вариантов, но обладают ключевыми ограничениями:

– низкая точность для редких мутаций;
– отсутствие специализированных моделей для RB1;
– слабая интерпретируемость предсказаний.

Цель исследования – разработка автоматизированной системы классификации патогенных мутаций RB1 на основе CNN.

Задачи исследования:

– создание сбалансированного датасета на основе ClinVar;
– построение архитектуры CNN с обработкой последовательностей ДНК;
– оценка эффективности модели (AUC ≥ 0.9, Precision ≥ 0.8).

Практическое применение – сокращение времени анализа мутаций с недель до часов при сохранении точности ≥ 85 %.

Инструментарий исследования – для данных 126 аннотированных мутаций RB1 (ClinVar), для модели CNN с Embedding-слоем и двумя сверточными блоками, TensorFlow, BioPython, scikit-learn, AUC-ROC, Precision-Recall, F1-score.

Искусственный интеллект (ИИ) и машинное обучение (МО) становятся незаменимыми инструментами в генетике и биоинформатике благодаря своей способности анализировать огромные объемы данных и выявлять сложные, неочевидные закономерности [3].

Разработанная нейронная сеть представляет собой многослойную архитектуру, специально адаптированную для обработки геномной информации. Она способна автоматически выявлять закономерности в мутационных данных, классифицировать их по степени опасности и предсказывать потенциальное клиническое значение новых мутаций.

Процесс создания нейронной сети для анализа мутаций гена RB1 состоит из нескольких этапов:

1. Сбор и предварительная обработка данных. Данные последовательности гена RB1 были загружены из FASTA-файла (RB1.fna), содержащего нуклеотидную последовательность RB1. Для обработки файла использовалась библиотека BioPython [4]. Информация о мутациях была взята из базы данных ClinVar в формате CSV (RB1_mutations.csv). В файле содержались данные о 126 мутациях, включая их геномные координаты, типы изменений и клиническую значимость. Из данных ClinVar были отобраны только мутации с маркерами Pathogenic или Likely pathogenic. Последовательности ДНК были преобразованы в числовой формат для дальнейшего анализа. Каждому нуклеотиду (A, T, G, C) был присвоен уникальный индекс. Для устранения дисбаланса классов был создан сбалансированный датасет, содержащий 160 примеров (по 80 для каждого класса).

2. Архитектура нейронной сети. Для анализа последовательностей ДНК была выбрана архитектура свёрточной нейронной сети (CNN), которая хорошо подходит для выявления локальных паттернов в данных [5]. Модель состояла из следующих слоёв:

– входной слой (принимал последовательности ДНК фиксированной длины (500 нуклеотидов));

– embedding-слой (преобразовывал индексы нуклеотидов в плотные векторные представления (размерность 8), что позволяло модели работать с семантически значимыми признаками);

– свёрточные блоки, состоящие из первого блока (32 фильтра с размером ядра 15, слой MaxPooling1D и Dropout (0.3). Этот блок был предназначен для выявления широких контекстов, таких как регуляторные области); второго блока (64 фильтра с размером ядра 7 и GlobalMaxPooling1D. Этот блок фокусировался на локальных паттернах, например, точечных мутациях); полносвязные слои, к которым относится Dense (32 нейрона) для сжатия признаков и Dense (1 нейрон с активацией sigmoid) для бинарной классификации (0 – норма, 1 – мутация).

3. Обучение модели. Обучение проводилось с использованием следующих параметров. Функция потерь – Binary cross-entropy, подходящая для задач бинарной классификации. Оптимизатор – Adam с начальной скоростью обучения 0.001. Метрики – Accuracy, Precision, Recall и AUC (Area Under Curve).Коллбэки – EarlyStopping (останавливал обучение, если качество на валидационной выборке не улучшалось). ReduceLROnPlateau (автоматически снижал скорость обучения при достижении плато).

4. Результаты обучения. На первых эпохах модель демонстрировала AUC 0.88 и Recall 1.0, что указывало на способность выявлять все мутации, но с некоторыми ложноположительными срабатываниями. К концу обучения точность на обучающей выборке достигла 85–90 %, однако на валидационной выборке наблюдалось переобучение, что подчеркивало необходимость увеличения объёма данных или применения дополнительных методов регуляризации.

Для интерпретации результатов использовались:

1. Графики метрик. Визуализация Accuracy, Loss, Precision и Recall на протяжении обучения (рис. 1).

Таким образом, точность на обучающей выборке постепенно растёт, достигая значений около 85–90 % к концу обучения. Однако точность на валидационной выборке остаётся низкой,

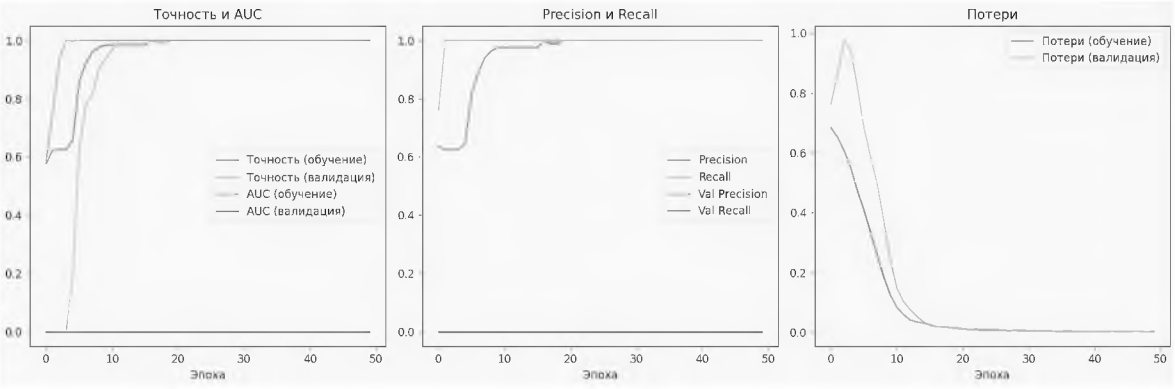

Рисунок 1 – Визуализация метрик обучения

что указывает на переобучение модели. Значение AUC улучшается на протяжении обучения, достигая уровня 0,92–0,95. Это свидетельствует о том, что модель способна эффективно отделять классы, даже если точность остаётся низкой. Precision остаётся на уровне 0,6, что указывает на большое количество ложноположительных предсказаний. Recall достигает значения 1,0, что означает, что модель успешно идентифицирует все примеры с мутациями, жертвуя точностью. Loss на обучающей выборке снижается, что указывает на успешное обучение модели. Однако валидационный loss остаётся высоким, что подтверждает необходимость увеличения объема данных или применения дополнительных методов регуляризации.

2. Анализ предсказаний. Изучение примеров, где модель допустила ошибки, для понимания её слабых мест (рис. 2).

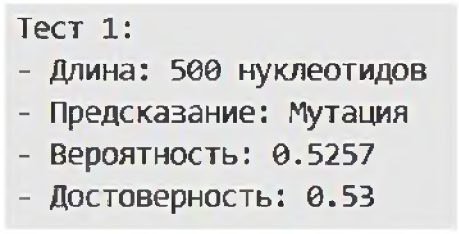

Рисунок 2 – Пример вывода

Модель демонстрирует способность к обучению, но страдает от переобучения из-за ограниченного размера датасета. Высокий recall при низком precision указывает на то, что модель склонна к предсказанию большинства примеров как мутаций. Это может быть связано с неоптимальным порогом классификации или недостаточной сложностью модели. Увеличение объема данных и применение методов регуляризации могут значительно улучшить качество модели.

Разработанная CNN показала потенциал для автоматизированного выявления патогенных мутаций в RB1, но требует доработок для повышения точности и устойчивости. Применение подобных моделей в клинической практике может ускорить анализ генетических данных и улучшить диагностику наследственных заболеваний, таких как ретинобластома.

Список использованных источников

1. Кнудсон, А. Г. Мутация и рак: статистическое исследование ретинобластомы / А. Г. Кнудсон // Труды Национальной академии наук США, 1971. – Т. 68, № 4. – С. 820–823.
2. ClinVar // Национальный центр биотехнологической информации США (NCBI). [Электронный ресурс]. – Режим доступа: URL: https://www.ncbi.nlm.nih.gov/clinvar/. – Дата доступа: 01.04.2025.
3. TensorFlow // Фреймворк для масштабируемого машинного обучения. [Электронный ресурс]. – Режим доступа: URL: https://www.tensorflow.org/. – Дата доступа: 27.03.2025.
4. BioPython // Библиотека для вычислительной молекулярной биологии. – Версия 1.81. – [Электронный ресурс]. – Режим доступа: URL: https://biopython.org/. – Дата доступа: 30.03.2025.
5. Педрегоса, Ф. Машинное обучение на Python: библиотека scikit-learn / F. Pedregosa, G. Varoquaux [и др.]; 2011. – Т. 12. – С. 2825–2830.

УДК 004.42+519.178

ВИЗУАЛИЗАЦИЯ ГРАФОВ НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ C++ С ПОМОЩЬЮ УТИЛИТЫ Graphviz

Гречаников А. А., студ., Соколова А. С., ст. преп.
Витебский государственный технологический университет,
г. Витебск, Республика Беларусь

Реферат. *В статье рассмотрена интеграция утилиты Graphviz в среду разработки Code::Blocks для визуализации графов на языке C++. Представлен практический пример использования алгоритма поиска минимального остовного дерева с автоматической генерацией графического представления.*

Ключевые слова: Graphviz, C++, минимальное остовное дерево, алгоритм Прима, Code::Blocks.