Терпіловський Єгор Олександрович, аспірант

Інститут кібернетики імені В.М. Глушкова НАН України, Київ

**Моделі та методи машинного навчання для оцінки ризиків генетичних захворювань**

Сучасний розвиток біоінформатики та машинного навчання відкриває нові можливості у сфері медичної діагностики, зокрема для визначення ризиків генетичних захворювань. Важливість дослідження генетичних ризиків обумовлена необхідністю їхнього раннього виявлення, що дозволяє підвищити ефективність лікування та профілактики. Генетичні захворювання часто мають складний механізм успадкування та залежать від взаємодії багатьох генетичних факторів. Аналіз ДНК дозволяє ідентифікувати характерні мутації, що можуть свідчити про наявність чи ймовірність розвитку захворювання. Проте традиційні методи аналізу генетичних даних є недостатньо точними або потребують значних обчислювальних ресурсів.

Метою дослідження є розробка та оптимізація моделей машинного навчання для класифікації ДНК з метою оцінки ризиків генетичних захворювань. Основна увага приділена методам обробки біологічних даних, виявленню характерних патернів у геномних послідовностях та адаптації моделей під реальні медичні потреби.

Серед основних завдань дослідження:

• Аналіз сучасних методів машинного навчання, що використовуються для класифікації ДНК.

• Розробка алгоритмів попередньої обробки ДНК-даних, що включає вирівнювання, виявлення SNP та нормалізацію послідовностей.

• Запропонування методів представлення ДНК-даних у вигляді векторів з урахуванням частотної та бінарної репрезентації k-мерів.

• Побудова класифікаційних моделей на основі ансамблевих алгоритмів (Random Forest, Gradient Boosting) та глибоких нейронних мереж.

• Використання адаптивних методів обробки дисбалансованих даних, що покращує якість прогнозування для вибірок із значною перевагою здорових або хворих зразків.

• Оптимізація параметрів моделей та тестування їх продуктивності на реальних даних секвенування.

Для досягнення поставлених завдань використовуються сучасні методи обробки та аналізу біологічних даних. Зокрема, впроваджено підхід на основі k-мерного аналізу, що дозволяє ідентифікувати специфічні патерни мутацій у ДНК та підвищити інформативність ознак. Для цього були проведені дослідження оптимальної довжини k-мерів, що забезпечує максимальну точність класифікації.

Застосовано комплексні методи попередньої обробки, що включають вирівнювання послідовностей до референсного геному, ідентифікацію однонуклеотидних варіацій (SNP) та нормалізацію даних. Розроблено алгоритми векторизації даних на основі частотного та бінарного представлення k-мерів, що дозволяє ефективно адаптувати дані для подальшої обробки нейромережами.

Створено класифікаційні моделі на основі ансамблевих методів (Random Forest, Gradient Boosting) та глибоких нейронних мереж. Для оптимізації моделей було проведено тестування різних гіперпараметрів, таких як кількість дерев у лісі, глибина дерев, швидкість навчання та кількість нейронів у прихованих шарах. Вперше впроваджено адаптивні алгоритми обробки дисбалансованих даних шляхом використання ресемплінгу та вагових коефіцієнтів у функціях втрат.

Запропонований підхід дозволяє значно підвищити точність класифікації ДНК до 95% та забезпечує ефективність аналізу навіть для великих обсягів біологічних даних. Проведене тестування моделей на відкритих наборах даних, таких як 1000 Genomes Project, підтвердило їхню стабільність та адаптивність. Розроблене програмне забезпечення інтегрує методи аналізу ДНК у автоматизовані діагностичні системи, що дозволяє масштабувати обробку даних та застосовувати методику в клінічних умовах.

Запропоновані моделі та методи відкривають нові можливості у діагностиці генетичних захворювань. Завдяки використанню машинного навчання вдалося підвищити точність аналізу та адаптувати методики під потреби персоналізованої медицини. Впроваджена система аналізу генетичних даних може бути застосована не лише у медичних дослідженнях, а й у біотехнологіях, фармакогенетиці та фундаментальній науці. Подальший розвиток дослідження спрямований на оптимізацію глибинних моделей, розширення наборів навчальних даних та вдосконалення методів виявлення мутацій. Це дозволить зробити методику ще більш точною, ефективною та доступною для широкого застосування.

**Література**

1. Y. Terpilovskyi, «The k-mer method in tasks of identifying regular sequences,» pp. 77-83, 2024.
2. Y. Terpilovskyi, «Comparative Analysis of DNA Classification Using the method of Random Forests and Convolutional Neural Networks,» pp. 96-103, 2024.
3. Y. Terpilovskyi, «Comparison of DNA k-mer data representations for classification via neural networks,» pp. 61-69, 2024.