

Пациент-ориентированное бенчмаркинг-исследование архитектур CNN и трансформеров для классификации гистопатологии рака молочной железы

Источник: Frontiers in Digital Health

Оригинал: <https://www.frontiersin.org/articles/10.3389/fdgth.2026.1752938>

CNN

бенчмаркинг

гистопатология

диагностика

компьютерное зрение

онкология

трансформеры

Введение

Диагностика рака молочной железы с использованием гистопатологических изображений остается критически важной, но сложной задачей, требующей автоматизированных систем, которые способны надежно обобщать данные на разных пациентах и при различных условиях визуализации. Несмотря на то, что модели глубокого обучения демонстрируют высокие показатели эффективности, во многих предыдущих исследованиях применялось разделение данных по отдельным изображениям, что приводит к утечке данных на уровне пациентов, в результате чего оценки становятся чрезмерно оптимистичными и потенциально вводящими в заблуждение. Данное исследование направлено на устранение этого ограничения путем создания строгой, свободной от утечек системы сравнительного тестирования (benchmarking framework) для бинарной классификации гистопатологии рака молочной железы.

Методы

Была проведена комплексная оценка девяти архитектур глубокого обучения на наборе данных **BreaKHis**, состоящем из 7 909 изображений 82 пациентов. Модели включают шесть сверточных нейронных сетей (**ResNet50**, **MobileNetV2**, **VGG16**, **DenseNet121**, **Xception** и **EfficientNetB0**), одну современную сверточную архитектуру (**ConvNeXt**) и две модели на основе трансформеров (**Swin-Small** и **Swin-Base**). Был внедрен строгий протокол 5-кратной перекрестной проверки с учетом данных пациентов (patient-aware cross-validation), чтобы гарантировать, что изображения одного и того же пациента не попадали одновременно в обучающую и валидационную выборки. Все модели обучались в идентичных экспериментальных условиях. Эффективность оценивалась с помощью показателей точности (**accuracy**), прецизионности (**precision**), полноты (**recall**) и **F1-score**, представленных как среднее значение \pm стандартное отклонение. Статистическая значимость оценивалась с использованием парного t-критерия Стьюдента и критерия знаковых рангов Уилкоксона с поправкой Бонферрони.

Результаты

Все оцененные архитектуры продемонстрировали сопоставимую эффективность, достигнув средних показателей точности в диапазоне 0,91–0,93. **ResNet50** достигла самой высокой средней точности ($0,9267 \pm 0,0435$) и показателя **F1-score** (0,9472), хотя различия между моделями были незначительными. Статистический анализ подтвердил, что никакие парные различия не были статистически значимыми ($p > 0,05$ после поправки). Анализ по степени увеличения показал, что промежуточные разрешения (40x и 200x) обеспечивают более дискриминативные признаки, в то время как более высокое увеличение (400x) приводило к снижению эффективности из-за ограниченной контекстуальной информации.

Обсуждение

Полученные результаты подчеркивают, что при использовании строго контролируемого протокола оценки, свободного от утечек данных, архитектурные различия между современными моделями глубокого обучения не приводят к статистически значимым вариациям производительности. Вместо этого, дизайн оценки играет более критическую роль в определении надежных результатов. Предложенная система сравнительного тестирования

с учетом данных пациентов повышает воспроизводимость и обеспечивает надежную основу для будущих исследований, способствуя разработке клинически применимых систем ИИ для диагностики рака молочной железы.

Перевод выполнен: 15.05.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.