Метод классификации лейкоцитов на изображениях клеток крови

Е. М. Черных, e-mail: jaddyroot@gmail.com
В. М. Михелев, e-mail: mikhelev@bsu.edu.ru

«Белгородский государственный национальный исследовательских университет» (НИУ «БеЛГУ»)

Аннотация. *В данной работе представлен метод классификации лейкоцитов на цифровых снимках клеток крови, полученных с помощью оптического микроскопа. Разработанный метод базируется на алгоритмах машинного обучения и включает в себя три этапа. Поочередное выполнение сегментации лейкоцитов, дополнительной обработки и обнаружения сегментированных объектов, а также заключительного этапа классификации распознанных клеток позволяет повысить точность результатов, полученных на предыдущем шаге. Приведенные результаты вычислительного эксперимента показали высокую точность работы предложенного метода, его эффективность и универсальность.*

Ключевые слова: *обработка изображений, распознавание образов. машинное обучение, сегментация, классификация лейкоцитов.*

Введение

На фоне существующих опасных диагнозов, а также растущей заболеваемости людей по причине охватившей весь мир пандемии, проблема своевременной постановки диагноза еще на начальной стадии развития заболевания стоит сейчас крайне остро. Одним из наиболее распространенных методов диагностики и оценки состояния здоровья человека является общеклинический анализ крови, а именно исследование состава лейкоцитов в пробе крови (лейкоцитарная формула). Эти белые клетки крови, являясь агентами иммунной системы человека, наиболее точно отражают информацию о любых проблемах и патологиях в организме [1]. Таким образом, вовремя обнаружив изменение уровня лейкоцитов определенного типа в крови, можно своевременно диагностировать заболевание, либо же назначить дополнительные исследования для уточнения деталей.

На протяжении нескольких десятков лет процесс подсчета и классификации лейкоцитов являлся кропотливой, преимущественно ручной работой, требующей больших временных затрат [2], а результат

такого анализа, выполняющегося специалистом с помощью микроскопа, получался сильно субъективным, поскольку зависел от знаний и опыта лаборанта. Однако в настоящее время в данной сфере получили распространение автоматизированные программно-аппаратные комплексы и системы, которые позволяют вести подсчет, анализировать и классифицировать широкий спектр форменных элементов крови на загружаемых в них мазках. К подобным системам можно отнести проточные цитометры, комплексы CellaVision Diff master Octavia и CellaVision DM96 [3-4], которые применяются в различных лабораториях по всему миру. Однако, данные комплексы имеют значительный недостаток – высокая стоимость оборудования и его сложность, требующая определенного уровня квалификации у оператора. Наряду с этим, с момента появления цифровой камеры, также неуклонно возрастало число попыток автоматизировать цитометрию на основе полученных цифровых изображений. Исследователи из разных стран предлагают медицинские решения на основе методов и алгоритмов компьютерного зрения совместно с машинным обучением, год за годом повышающих свою надежность в решении задач анализа изображений. Но, к сожалению, значительная часть данных решений и методов непригодны для реальной медицины, так как их разработка и тестирование несли скорее научный характер, нежели пригодный для практического применения в сфере здравоохранения.

Исходя из этого, можно сделать вывод о том, что доступный эффективный метод решения задачи подсчета и классификации лейкоцитов на цифровых изображениях, который бы позволял быстро и качественно получать результаты анализа снимков, обеспечивая достаточную для практической медицины точность работы, до сих пор отсутствует, и его разработка сейчас – одна из первостепенных и актуальных задач клинической медицины.

Таким образом, в рамках данной работы, объектом проведенного исследования выступали цифровые цветные изображения клеток крови, а предметом исследования – метод классификации лейкоцитов на данных изображениях.

1. Теоретические основы предлагаемого метода

В ходе проведенного анализа подходов к решению данной проблемы, было установлено, что современные исследователи подходят к решению задачи классификации лейкоцитов на изображениях в несколько этапов, среди которых наиболее часто встречаются: предварительная обработка изображения; сегментация и/или обнаружение лейкоцитов; извлечение признаков из обнаруженных

клеток; этап классификации клеток. Предлагаемый метод состоит из трех этапов работы с исходным изображением:

Сегментация, цель которой – выделить области, содержащие лейкоциты, из исходного изображения относительно общего фона и других форменных элементов на снимке. В ходе проведения анализа по поиску метода сегментации лейкоцитов, который бы обеспечивал наибольшую эффективность и универсальность, было принято решение использовать в сверточную нейронную сеть архитектуры U-net, повысив ее начальную точность за счет трансферного обучения – техники, которая предоставляет возможность применить уже обученную для решения одной задачи модель для решения другой целевой задачи. В данной работе предлагается использовать трансферное обучение для того, чтобы обучить U-net на основе больших и сложных архитектур, которые были предварительно обучены на большом наборе данных ImageNet, состоящим из более 14 миллионов изображений нескольких тысяч различных классов [5]. В качестве предобученных моделей в данной работе используются архитектуры InceptionResNetV2 и две разные модели семейства EfficientNet. На рис. 1 приведены результаты сегментации для нескольких различных изображений. Для повышения конечное точности сегментации было предложено объединить три данные модели в ансамбль. В данной работе используется создание ансамбля на основе именно взвешенного голосования: прогноз каждой модели умножается на вес, после чего вычисляется их среднее значение, которое и является конечным результатом работы ансамбля [6]. Результатом этапа сегментации является черно-белое изображение (бинарная маска), где белые пиксели характеризуют области с лейкоцитами, а черные – все прочие пиксели. В большинстве случае, дальнейшая работа с необработанной бинарной маской может привести к плохим результатам классификации по причине наличия артефактов и ложно обнаруженных объектов.

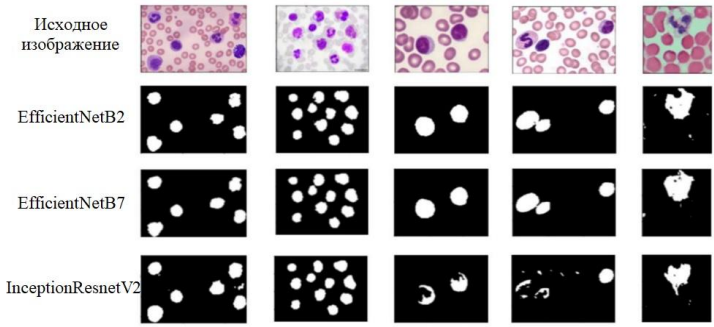


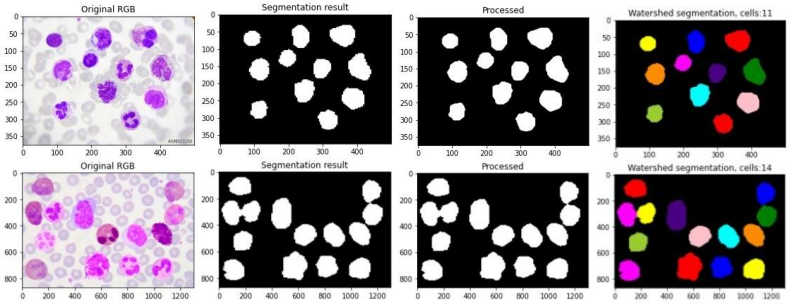
Рис. 1. Примеры результатов сегментации различными моделями

Дополнительная обработка результатов сегментации, **обнаружение** и **извлечение** сегментированных лейкоцитов является следующим модулем, который отвечает за обработку полученной в результате сегментации бинарной маски. Данная обработка направлена на устранение двух основных проблем: очистки результатов сегментации от артефактов, лишних пикселей, ложно сегментированных объектов и т. д., что достигается путем:

1. *математическая морфология*: операции замыкания и размыкания. Очищает изображение от лишних шумов и артефактов, сглаживает и незначительно увеличивает расширяет границы сегментированных объектов;
2. *расчет площадей объектов* на изображении. Выполняется маркировка связных областей пикселей (смежных, или имеющих одинаковое значение), а затем вычисляется площадь каждой маркированного области.
3. *очистка изображения* от объектов, площадь которых значительно меньше средней площади. На основе медианного значения среди всех полученных на предыдущем шаге площадей связных областей вычисляется средняя площадь объектов, поскольку, предполагается, что масштаб клеток может быть разным, но их площади лежат в одном среднем диапазоне. Очистка маркированных областей, которые значительно больше или меньше средней площади предполагаемых клеток, позволяет очистить бинарную маску от лишних объектов;
4. *применение алгоритма Watershed* для выделения связанных компонентов в отдельные объекты с учетом склеенных или наложенных друг на друга клеток;

5. *подсчет оставшихся объектов*, лежащих в пределах средней площади объектов;
6. *отрисовка контуров* вокруг объектов с небольшим отступом;
7. *извлечение объектов* из исходного цветного изображения на основе отрисованных контуров.

На рис. 2 показаны примеры промежуточных результатов модуля дополнительной обработки, среди которых можно увидеть, как выполняется сегментация оригинального изображения, как происходит очистка результатов сегментации, после чего применяется алгоритм Watershed для обнаружения и подсчета лейкоцитов на изображении с учетом того, что клетки могут быть склеены и/или перекрыты друг другом.



а б в г

а – исходные изображения, б – результат сегментации, в – результат дополнительной обработки, г – результат применения алгоритма Watershed

Рис. 2. Промежуточные результаты модуля дополнительной обработки

Классификация, поочередно выполняющаяся для извлеченных из исходного изображения областей с лейкоцитами относительно пяти основных классов этих клеток. Как и для решения задачи сегментации, при обучении сверточной нейронной сети использовалась техника Transfer Learning, в качестве предобученной модели в модуле классификации используется также упомянутая ранее архитектура InceptionResNetV2 [7], как показавшая наилучшие результаты работы модель. На рис. 3 изображены примеры результатов классификации извлеченных областей с лейкоцитами, где в верхней строке над

изображением указан истинный класс клетки, а в нижней строке – класс, предсказанный моделью.

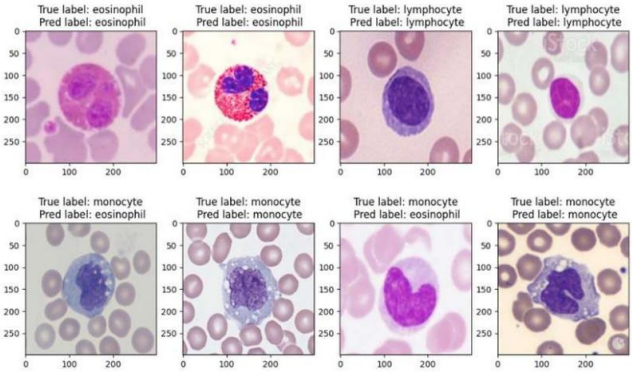


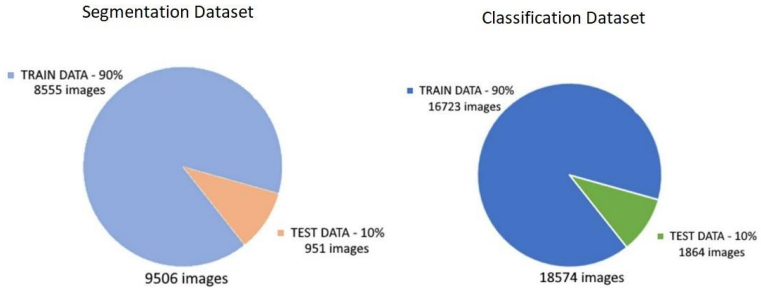
Рис. 3. Пример результатов классификации

2. Исходные данные и их структура

Поскольку основная часть предлагаемого метода базируется на моделях сверточных нейронных сетей, для разработки метода предварительно потребуется выполнить обучение данных моделей. Первый модуль метода, отвечающий за сегментацию лейкоцитов, использует ансамбль трех объединенных вместе различных моделей, а третий модуль, выполняющий роль классификатора клеток, использует одну обученную модель. Обучение данных четырех моделей производилось раздельно, на различных наборах данных. Сравнивая результаты обучения моделей для сегментации и классификации, было обнаружено, что и в обоих случаях использование более, чем одного набора данных для обучения модели позволяет сделать ее более устойчивой к данным нового типа, что, соответственно, повышает ее конечную практическую значимость. По этой причине для обучения сегментационных моделей был использован составной датасет, состоящий из трех различных наборов данных. Аналогичным образом строится и второй датасет, для обучения модели решать задачу классификации: он сформирован из трех различных наборов данных.

Круговые диаграммы на рис. 4 отражают состав и структуру датасетов, которые использовались для обучения сегментационных моделей и модели-классификатора. Общее количество пар изображений, использовавшихся для обучения моделей сегментации – более 9500. 90% от этого объема было использовано для обучения моделей, а 10% – для оценки их точности. В этот составной набор вошли изображения из

датасетов: от Jiangxi Tecom Science Corporation (Китай) и CellaVision [8], а также известная база данных LISC, собранная специалистами центра гематологии в Иране [9]. В состав второго датасета вошли следующие наборы данных: часть упомянутого ранее набора единичный клеток [8], но без использования бинарных масок, набор данных из репозитория Mendeley [9], а также классифицированные изображения лейкоцитов, распространяемые на открытом ресурсе Kaggle [10]. Более 18 тысяч картинок аналогичным образом разбиваются на 90% и 10% в качестве обучающей и тестовой подвыборок.



а б

а – набор данных для сегментации, б – набор данных для классификации

Рис. 4. Структура обучающих наборов данных

3. Вычислительный эксперимент

Ввиду того, что предлагаемый метод в большинстве своем базируется на работе обучаемых моделей искусственных нейронных сетей, для оценки точности данных моделей используется специальный подготовленный тестовый набор данных, который упоминался в предыдущем разделе. Важным условием формирования тестового набора данных является то, что входящие в него изображения не должны быть использованы для обучения модели. Для оценки точности трех обученных сегментационных моделей было использовано 950 тестовых изображений, для которых модели выполняли сегментацию, и после чего полученные результаты сравнивались с истинными данными. В таблице ниже приведены полученные значения основных метрик, использующихся при оценке точности моделей: индекс Жаккара, используемый в качестве основного показателя точности при решении задачи сегментации изображений [11], точность, полнота и F-мера. В нижней строке таблицы можно увидеть также повышение значений данных метрик при объединении трех моделей в ансамбль.

Таблица

Показатели основных метрик обученных моделей при их оценке на тестовых данных

Модель	IoU, %	Accuracy, %	Recall, %	Precision, %	F1, %
EfficientnetB2	99.21	99.82	99.21	99.40	99.30
EfficientnetB7	99.41	99.87	99.55	99.42	99.48
InceptionResnetV2	99.22	99.82	99.34	99.28	99.31
Ансамбль моделей	99.48	99.88	99.57	99.52	99.54

На рис. 5 представлена матрица ошибок (confusion matrix), рассчитанная для ансамбля моделей.

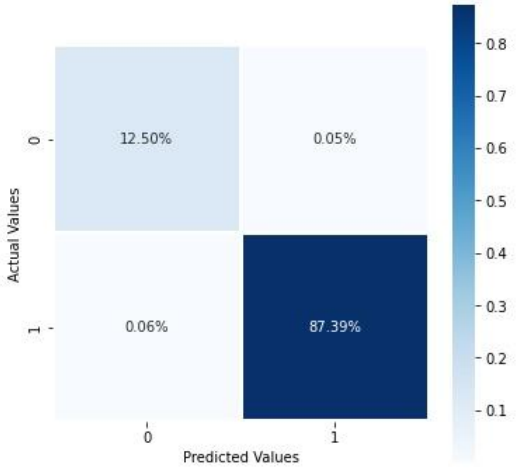


Рис. 5. Матрица ошибок (confusion matrix) для ансамбля моделей, решающего задачу сегментации

При оценке точности работы классификационной модели была также использована отдельная тестовая выборка, содержащая более 1800 изображений и соответствующих им меток классов. На рис. 6 представлена полученная матрица ошибок для классификатора, которая имеет большее число строк и столбцов, поскольку количество возможных классов для данной задачи равно 5, в отличие от задачи

сегментации, которая по своей сути является в данном случае бинарной. Итоговая точность классификационной модели составила 92.69%

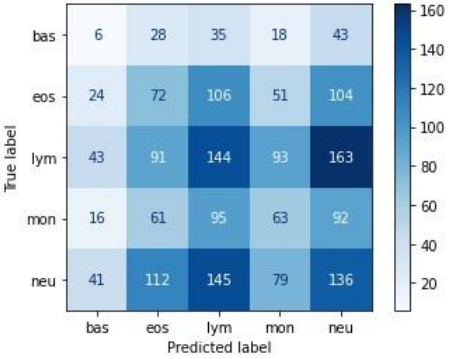


Рис. 6. Матрица ошибок для модели-классификатора

Заключение

Предложенный в данном исследовании метод классификации лейкоцитов на цифровых снимках форменных элементов крови позволяет упростить и ускорить процесс решения данной задачи за счет использования таких современных технологий, как алгоритмы компьютерного зрения и машинного обучения. Полученные в ходе проведения вычислительного эксперимента результаты тестирования показали, что метод обладает высокой точностью работы на каждом из трех этапов по отдельности, а также при их совместной работе в составе общего метода. Также предложенный метод нивелирует один из самых значительных недостатков существующих подходов – он обладает способностью к обобщению, являясь универсальным.

Данный метод классификации обладает высокой практической значимостью, поскольку может быть применен на практике в гематологических лабораториях как вспомогательный инструмент для определения лейкоцитарной формулы и подсчета лейкоцитов на цифровых изображениях.

Список литературы

1. Смирнова О. В., Савченко А. А., Манчук В. Т. Иммунометаболические механизмы развития острых лейкозов. – 2011.
2. Черных Е. М., Михелев В. М. Компьютерная система классификации лейкоцитов на изображениях клеток крови //Научный результат. Информационные технологии. – 2019. – Т. 4. – №. 3. – С. 38-47.
3. Kratz A. et al. Performance evaluation of the CellaVision DM96 system: WBC differentials by automated digital image analysis supported by an artificial neural network //American journal of clinical pathology. – 2005. – Т. 124. – №. 5. – С. 770-781.
4. Swolin B. et al. Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks–evaluation of DiffMasterTM Octavia //clinical & laboratory haematology. – 2003. – Т. 25. – №. 3. – С. 139-147.
5. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks //Advances in neural information processing systems. – 2012. – Т. 25. – С. 1097-1105.
6. Kuncheva L. I., Rodríguez J. J. A weighted voting framework for classifiers ensembles //Knowledge and Information Systems. – 2014. – Т. 38. – №. 2. – С. 259-275.
7. Szegedy C. et al. Inception-v4, inception-resnet and the impact of residual connections on learning //Proceedings of the AAAI Conference on Artificial Intelligence. – 2017. – Т. 31. – №. 1.
8. Zheng X. et al. Fast and robust segmentation of white blood cell images by self-supervised learning //Micron. – 2018. – Т. 107. – С. 55-71.
9. Acevedo A. et al. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems //Data in Brief, ISSN: 23523409, Vol. 30, (2020). – 2020.
10. Matek C. et al. A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls (AML-Cytomorphology_LMU) //The Cancer Imaging Archive (TCIA). – 2021.
11. Zheng Z. et al. Distance-IoU loss: Faster and better learning for bounding box regression //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 07. – С. 12993-13000.