МОРФОЛОГИЧЕСКАЯ ДИАГНОСТИКА КЛЕТОК РИД — ШТЕРНБЕРГА ПРИ ЛИМФОМЕ ХОДЖКИНА С ПОМОЩЬЮ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

В.В. Пунько[1], Н.А. Волорова[1], В.С. Приходько[2]

[1]Белорусский государственный университет информатики и радиоэлектроники

[2]Гродненский государственный медицинский университет

Аннотация. Предложен способ морфологической диагностики клеток Рид – Штернберга при лимфоме Ходжкина с помощью искусственных нейронных сетей. Разработана архитектура комплекса программных средств для анализа препаратов (биопсий), состоящего из модулей для работы с данными и искусственными нейронными сетями. Рассмотрены способы обучения нейронных сетей в условиях нехватки данных для рассматриваемой и подобным им задачам.

Ключевые слова: искусственные нейронные сети, диагностика онкологических заболеваний, обучение искусственных нейронных сетей на малых выборках данных.

Abstract. Suggested a way to morphologically diagnose Reed – Sternberg cells in case of Hodgkin's lymphoma using artificial neural networks. Designed complex program architecture consisting of blocks for processing data and artificial neural networks for tissue sample analysis. Researched problem of training neural networks with lack of representative datasets for this task and other similar ones.

Keywords: artificial neural networks, cancer diagnosis, training artificial neural networks on small datasets.

Введение

Онкологическое заболевание – одно из самых распространенных трудно диагностируемых заболеваний, которые всё чаще и чаще встречаются среди населения, и это никак не зависит от экономического статуса или социального положения людей. Этот тип заболеваний одни из самых летальных и проигрывают в смертности только сердечно-сосудистой патологии. Несмотря на весь прогресс в диагностике и методах лечения темпы увеличения количества онкобольных неуклонно растут, и проблема онкологии на данный момент является одной из самых актуальных в современной медицине.

Различных онкологических патологий огромное количество. Одного только рака молочной железы существуют десятки видов. Кроме того, у каждого вида есть различные подвиды, которые могут влиять на процесс и методы лечения. К примеру, самым распространенным видом онкологического заболевания являются рак легких, груди, кишечника и простаты, которые составляют порядка 45% всех случаев заболевания.

Несмотря на то, что онкологическая диагностика является актуальной темой, про неё очень мало знают вне профессиональной среды. От этапа попадания пациента к онкологу до этапа, когда онколог подозревает у пациента профильное заболевание и ставит диагноз, проходит много шагов. Эти шаги и называются онкологической диагностикой.

Если в результате обследования обнаружено какое-либо новообразование, то это еще не позволяет делать выводы о том, что у пациента рак. Роль диагностики в онкологии очень высока, и при наличии подозрения на новообразование, пациент отправляется на диагностики, который называется морфологическим обследованием.

Морфологическое обследование подразумевает забор материала (биопсия) и собственно морфологическое исследование, на основании которого делается заключение о диагнозе. От точности диагностики на этом этапе будет зависеть, подойдёт ли назначенное лечение пациенту и насколько оно будет эффективно. Однако для того, чтобы говорить о диагностике какого-либо заболевания, нужно представлять его сущность.

В данной работе рассматриваются методы диагностики одного из онкологических заболеваний, а именно лимфомы Ходжкина или лимфогранулематоза. Лимфогранулематоз [1] – злокачественная патология, характерным признаком которой является наличие огромных клеток Рид – Штернберга. Он обнаруживается при микроскопическом исследовании и является одним из самых распространенных видов рака на сегодняшний момент (порядка 14% всех случаев). По всему организму рассеяна гигантская сеть лимфатических узлов, связанных тонкими сосудами. Эта сеть в организме занимается фильтрацией, и поэтому через нее постоянно проходят токсины и прочие вредные вещества, которые могут вызвать мутации. При неблагоприятных условиях или наличии мутаций нарушается процесс образования новых лимфоцитов, некоторые формы которых и составляют новообразование.

Данный вид рака можно обнаружить только морфологическим анализом. Для этого у пациента берут биопсию и подготовив препарат смотрят в микроскоп и визуально оценивают воз-

можность - данный анализ может быть неверным в силу ряда причин, основной из которых является человеческий фактор. В данной работе предлагаются методы повышения точности и эффективности проведения этого этапа, и удаления фактора человеческой ошибки из анализа путем создания комплекса программных средств для анализа биопсии.

Описание задачи и теоретические сведения

Исходной информацией для проведения анализа являются изображения биопсий, на которых видны межклеточное вещество и различные клетки, среди которых нужно найти больные, если таковые имеются. Цель этой работы – разработка комплекса программных средств для обнаружения больных клеток. Данную задачу морфологического анализа можно разбить на три основных этапа. На первом этапе отделяются клетки от межклеточного вещества. На втором они сортируются на здоровые клетки и возможно больные (обычно б**о**льшая часть клеток здоровые и такая сортировка очень сильно ускоряет работу). На третьем отобранные клетки классифицируются по заболеваниям.

Использование обычных детерминированных алгоритмов будут не эффективным так как данная задача является достаточно специфическое и для принятия решения следует учитывать огромное число факторов. Каждая новая биопсия является уникальной, потому здесь нужен интеллектуальный алгоритм, который сможет давать ответ опираясь на опыт предыдущих результатов. Поэтому было решено использовать искусственные нейронные сети.

Искусственная нейронная сеть – это математическая модель, которая имитирует биологические нейронные сети. Она представляет собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов), каждый из которых имеет дело только с сигналами, которые он получает от других процессоров.

В большинстве случаев нейронные сети учатся алгоритмом, который называется обучение с учителем. Для использования этого алгоритма требуется наличие репрезентативной выборки данных, специально подготовленной для этого процесса (для каждого элемента из этой выборки известен результат).

Проблема в том, что выборок для нашей задачи недостаточно для использования аппарата нейронных сетей, и приходится работать с той выборкой, которая, имеется в нашем распоряжении. Для искусственного увеличения данных были использованы два метода. Первый метод увеличения количества данных, который был применён, – это кросс-валидация по k-блокам [2] (при обычном обучении выборка обычно разделяется на два блока: на одном нейронная сеть обучается, а на втором она тестируется; при кросс-валидации она обучается на каждом из блоков и тестируется на противоположном, что, по сути, позволяет удвоить количество данных для обучения). Второй – аугментация (искусственное изменение данных). Эти методы реализованы и были применены в нескольких блоках нашего комплекса.

С учетом того, что сеть глубока, а имеющаяся выборка небольшая по объем, то возникла проблема переобучения суть которой заключается в том, что сеть запоминает исходную выборку и не учится на среднее. Она решалась с помощью добавления дополнительных промежуточных слоев в сети между уже имеющимися под названиями dropout и batch-normalize, а также с помощью регуляризации весов [3] и выбивания нейронов.

Помимо этого, для смягчения последствий двух вышеуказанных проблем был использован автокодировщик [4] – это однослойная нейронная сеть, предназначенная для уменьшения шума данных, которая, как побочный эффект, еще и всегда выдает разный результат даже на одном и том же входном значении, что уменьшает вероятность переобучения и помогает слегка увеличить размер выборки.

Архитектура программы

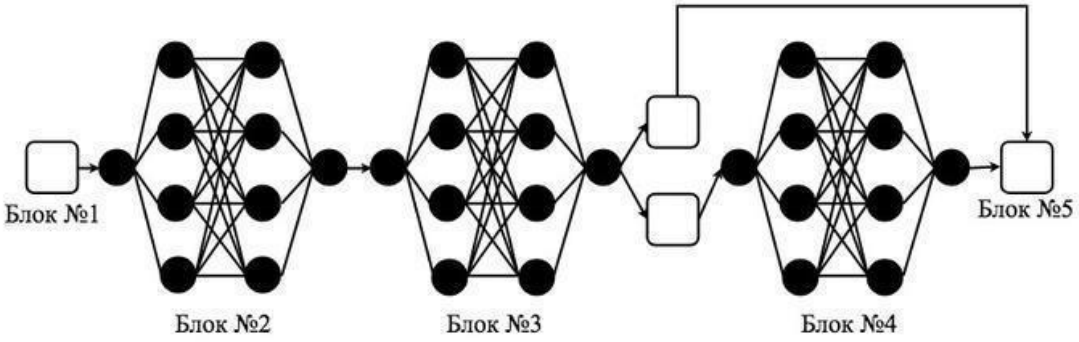

Рисунок 1. Схема архитектуры комплекса программных средств

Блок №1. Первым блоком в разработанной архитектуре (Рисунок 1) является препроцессинг. На данном этапе изображение проходит предварительную подготовку для работы с нейронной сетью. Этот этап очень важен, потому что изображения отличаются по цвету, размеру и наличию шумов от одной биопсии к другой. Препроцессинг унифицирует вид изображений так, чтобы нейронная сеть могла с ними стабильно работать. Для этого выполняются следующие действия:

1. Изображение переводится из цветовой модели RGB в цветовую модель HSV, так как в RGB нет возможности работать с яркостью пикселов.

2. Яркость полученного изображения повышается до максимума, для того, чтобы убрать различия, которые были созданы на препаратах (биопсий) из-за разницы в подсветках микроскопов.

3. Изображение возвращается в цветовую модель RGB и переводится в черно-белый формат, так как таким образом сегментировать изображение будет намного проще.

4. Для повышения контрастности изображения используем эквализация гистограммы изображения (Рисунок 2).

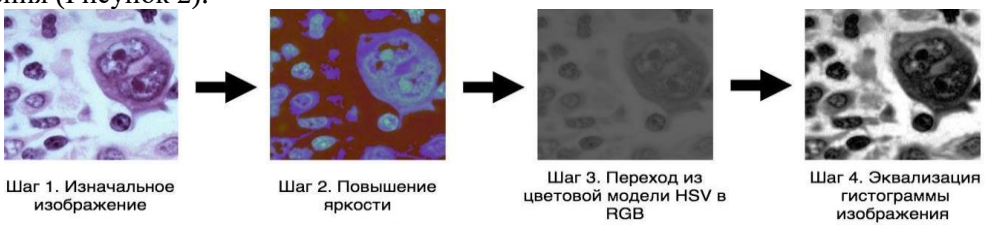

Рисунок 2. Пример преобразования изображение в препроцессинге

Блок №2. Вторым блоком является искусственная нейронная сеть с архитектурой, предназначенной для сегментирования изображений. На этом этапе все клетки, которые присутствуют на препарате (биопсии), должны быть отделены от межклеточного вещества. Используется архитектура нейронной сети U-Net: Convolution Network for Images Segmentation [5]. Данный тип нейронных сетей способен “накладывать” маску на изображение, тем самым убирая ненужные элементы, что достигается за счет свертки исходного изображения до вектора признаков (этот этап называется convolution). После чего, применяя полученные в ходе свертки признаки, изображение разворачивается до исходного (этот этап называется deconvolution), тем самым убирая лишние элементы.

В результате этой процедуры переданное с препроцессинга изображение будет обработано сетью и сеть создаст черно-белую маску для обрезания, где чёрный цвет – это межклеточное вещество, а белый – сами клетки. Далее по этой маске обрезается изначальное изображение (не прошедшее препроцессинг), причём оно обрезается таким образом, чтобы на каждом полученном изображении была только сама клетка. Все полученные изображения помещаются в вектор, который передается на следующий блок.

Блок №3. Третьим блоком является сверточная нейронная сеть, которая служит для сортировки клеток на те, которые точно здоровы, и те, которые возможно больны. Значительная часть клеток (около 80%), по статистике, на препарате являются здоровыми и будут исключены. Это помогает как ускорить процесс работы нейронной сети, так и сделать следующую нейронную сеть тоньше, так как для обучения тонкой нейронной сети требуется меньше данных и времени.

Сортировка клеток – это одна из задач классификации. Для этого используется искусственная нейронная сеть convolution neural network (CNN) с модификацией Xception [6], которая

позволяет сделать компактную глубокую сеть, нежели использование обычного алгоритма convolution neural network без модификаций. Данная модификация позволяет обрабатывать пространственную информацию и межканальную информацию изображения последовательно без потери качества работы сети и раскладывает обычную свертку на два этапа - pointwise convolution (обработка межканальной корреляции) и spatial convolution (обработка пространственной корреляции). Такие сети имеют меньше весов, а модель не проигрывает в точности.

В результате этого шага, все клетки, полученные путем сегментации изображения будут отсортированы на два класса – здоровые и возможно больные.

Блок №4. На этом этапе возможно больные клетки проходят процесс классификации, который очень схож с тем, который проводился в третьем блоке, однако классифицируются уже больные клетки по заболеваниям. Теоретически, данную модель можно натренировать на любое количество заболеваний, но в данном работе рассматривается лимфома Ходжкина. Рассматриваемая сеть имеет два класса и псевдокласс (клетки которые были классифицированы как возможно больные, но не являются таковыми): первый класс – лимфома Ходжкина, второй – все остальные больные клетки, класс заболевания которых нас не интересует. После этого сеть подает данные на последний блок приложения.

Блок №5. Последний блок формирует отчет о всех клетках на препарате, сколько здоровых, сколько больных и главное, где находятся раковые клетки. Это возможно сделать, имея данные, которые получил сегментатор на втором шаге нашего приложения. Отчет может использовать пациент для консультации с врачом или сам врач, ставя диагноз пациенту.

Заключение

В ходе данной работы была разработана архитектура приложение, которая способна диагностировать рак по биопсии, что облегчит работу врачей, и, в дальнейшем, поможет диагностировать раковые заболевания на ранних стадиях, тем самым спасая жизни людей. Также были решены некоторые проблемы с нехваткой данных и собрана репрезентативная выборка лимфогранулематоза, предоставляя возможность дальнейшего изучения этой разновидности рака методом статистического анализа.

Список литературы

1. Лимфогранулематоз / Л. П. Симберцова, Л. Холсти / М.: Медицина, 1985 - 304с.
2. Машинное обучение на Python / Francois Chollet / СПб.: Питер, 2018 - 400с.
3. Deep learning / Ian Goodfellow, Yoshua Bengio, Aaron Courville / MA: MIT Press, 2017 - 800c.
4. Gradient-based learning applied to document recognition [Электронный ресурс]. - URL: http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf (дата обращения: 06.04.2018).
5. U-Net: Convolutional Networks for Biomedical Image Segmentation [Электронный ресурс]. - URL: https://arxiv.org/pdf/1505.04597.pdf (дата обращения: 25.07.2018).

Xception: Deep Learning with Depth Wise Separable Convolutions [Электронный ресурс]. - URL: https://arxiv.org/pdf/1610.02357.pdf (дата обращения: 23.08.2018).