Проект 5

Реализация ансамбля алгоритмов машинного обучения для дифференциации кожных патологий на спектрах комбинационного рассеяния

|  |  |
| --- | --- |
| Фамилия, имя, отчество (полностью) | Сержантов Кирилл |
| Место учебы/работы (полностью) | ФГБОУ ВО «Тольяттинский государственный университет» |
| Номинация конкурса | Инжиниринговые проекты |

**СВЕДЕНИЯ О ПРОЕКТЕ**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название проекта | | Реализация  ансамбля алгоритмов машинного обучения для дифференциации кожных патологий  на спектрах комбинационного рассеяния | | |
| Аннотация проекта | | Рассматривается проблема неинвазивной диагностики онкологических заболеваний тканей кожи человека. Проводится предобработка результатов спектроскопии комбинационного рассеивания, с целью устранения шумов присутствующих в необработанных спектрах, для составления наборов данных пригодных при обучении алгоритмов машинного обучения.  Рассматривались спектры комбинационного рассеивания. Предобработка осуществлялась в три приема, выделение информативной части спектра 803.01-945 нм. сглаживание при помощи фильтра Савицкого-Голея, стандартизация полученных результатов. Работа осуществляется на языке высоко уровня Python 3.8, с использованием библиотек NumPy, Pandas.  Для дифференциации спектров используются методы ансамблирования: голосования большинства и стекинг реализованный на основе алгоритмов машинного обучения опорные вектора, логическая регрессия, k ближайших соседей, случайный лес. По обученной выборке была достигнута точность дифференциации патологии с точностью 90%. | | |
| Сроки реализации проекта | | Дата старта | Дата окончания | |
| 31.03.2020 | 31.03.2021 | |
| География проекта | | Самарский областной клиническийонкологический диспансер(ГБУЗ СОКОД).  Предварительно рассматриваются онкологические отделения диспансеров и поликлиник Самарской области. | | |
| Целевая аудитория | | Онкологические отделения диспансеров и поликлиник | | |
| Решаемая проблема | | На сегодняшний день в России сохраняется тенденция по заболеваемости онкологическими патологиями. Порядка 330 случаев на сто тысяч населения. И рак кожи занимает первые места по диагностируемости. Приведем статистику за 2018 год. 78,5 тысяч заболевших раком кожи из которых 5,46 тысяч приходится на меланому. При данном диагнозе процент летальных исходов составляет порядка 75%. Основной причиной является позднее диагностирование.  Трудности, при определении меланомы врачами общей практики, связаны со сложностью в интерпретации клинических признаков опухоли и невозможностью отличить меланому от доброкачественных пигментных образований (таких, как невус) на ранней стадии развития.  Дополнительной трудностью является отсутствие возможности использовать инвазивные методы исследования патологии, таких как биопсия с гистологическим или цитологическим исследованием, из-за повышенного риска прогрессирования опухоли.  В связи с этим, оптические методы имеют огромный потенциал для неинвазивного выявления и распознавания типа новообразований в тканях кожи.  Спектры комбинационного рассеяния сложны. Традиционное распознавание с помощью визуального осмотра является субъективным методом, отнимающим много времени.  Использование алгоритмов машинного обучения для решения этой задачи позволит нам создать модель, способную находить скрытые зависимости и извлекать шаблоны из необработанных данных, которые отвечают за классификацию злокачественных новообразований кожи, повышая точность распознавания патологии и в перспективе способную регистрировать смертельные новообразования на ранних стадиях, помогая врачам своевременно оказывать квалифицированную помощь. | | |
| Значимость проекта | | Разработка и внедрение современных технологий обработки спектров комбинационного рассеяния (КР) и автофлуоресценции(АФ) увеличивают шанс обнаружить болезнь на ранней стадии и не допустить ее дальнейшего прогрессирования, значительно облегчая лечение больного и повышая тем самым показатели здоровья населения. | | |
| Цель проекта | | Целью проекта является разработка ПО для неинвазивногообнаружения патологий тканей кожи на ранних стадиях развития заболевания. | | |
| Задачи | | 1. предобработать спектры АФ и КР. 2. Отобрать классификаторы машинного обучения для такого типа данных. 3. Обучить и протестировать классификаторы. 4. Проанализировать полученные результаты. 5. Выполнить дифференциацию патологии среди патологий, например для меланомы, как для самого агрессивного типа злокачественного новообразования. 6. разработать ПО, для установки в онкологических отделениях поликлинник. | | |
| Результаты и способ их измерения | | Для оценки работы нейронной сетииспользуется F-мера. F-мера объединяет в себе 2 метрики: точность, показывающую как много снимков отмеченных как патология действительно оказались с патологией, и полноту, показывающую как часто нейросеть ошибается на снимке без патологии. Также F-мера позволяет дать одной из метрик больший вес в итоговом результате. В задаче двоичной классификации ансамбль алгоритмов машинного обучения позволил нам получить значения: f1 - 90%, специфичность - 93%, чувствительность - 88%, доверительный интервал ансамбля 67,5% - 83,5%.Полученная точность сопоставима с большинством исследований с точностью 80% -90%.  Но наше основное отличие от большинства работ заключается в преобладании образцов меланомы кожи в наборе тренировочных данных. | | |
| Методы реализации проекта | | В ходе научно исследовательской работы был разработан модуль для первичной обработки входных данных. Программный код был написан на языке высокого уровня Pathon, версии 3.8 с использованием библиотек numpy, pandas.  Разработанный программный модуль предполагает возможность его использования как на новых данных, так и для первичной обработки датасета пригодного для последующего обучения алгоритмов машинного обучения. Листинг кода представлен в приложении A.  Предобработка была разделена на 3 шага:   * выделение информативной части спектра 803.01-945 нм; * сглаживание фильтром Савицкого; * стандартизация спектров.   Был выполнен PCA анализ, для уменьшения количества входных данных спектра при его анализе. Для определения оптимального значения параметров выборки использовалась функция поиска по сетке. Полученные параметры были проверены по алгоритму логистической регрессии. Мы использовалиGridSearchCV, чтобы установить размер PCA. Таким образом, все значения проверялись и выбирались оптимальные для данного образца. В нашем случае для каждого образца было выбрано 5 параметров.  Полученными данными был обучен ансамбль классификаторов машинного обучения. Ансамбль алгоритмов машинного обучения позволил нам получить значения: f1 - 90%, специфичность - 93%, чувствительность - 88%, доверительный интервал ансамбля 67,5% - 83,5%. | | |
| Возможность коммерциализации проекта | | Программное обеспечениебудетпредоставляться в онкологические отделенияполиклиник как дополнительный инструмент («второй глаз») врача для обнаружения патологий на спектрах КР и АФ. | | |
| Привлечение общественности и профессионального сообщества к реализации проекта | | Информационную поддержку в работе над проектом предоставляют врачи ГБУЗ СОКОД, консультируя по всемвопросам, а также предоставляют для обучения алгоритмов машинного обучения выборку с таблицей принадлежности спектров к патологиям. | | |
| Информация о команде проекта | | Лисовская Мария Германовна –старший преподаватель кафедры "Прикладная математика и информатика" ТГУ,  Сержантов Кирилл Александрович- магистрант ТГУ  Мякинин Олег Олегович – старший преподаватель кафедры «Лазерные и биотехнические системы» СНИУ им. С.П. Королева. | | |
| Личный вклад каждого члена команды в разработку проекта | | Олег Олегович и Мария Германовна, как руководители проекта курирует всю работу, обеспечивают связь с больницей и предоставляют необходимые данные, а также исходя из своего опыта и знаний направляют в разработке.  Кирилл Александрович проводит анализ полученных данных и разрабатывает приложение для решения поставленных задач. | | |
| Информационное сопровождение проекта | | Спектры КР И АФ были получены и проведены консультации у онкологов.  Данные спектры были проанализированы, сегментированы, выделена информативная область.Была проанализирована научная литература по решению задачи такого типа. Отобраны классификаторы машинного обучения. Протестированы на полученных данных. | | |
| Устойчивость проекта (дальнейшее развитие проекта) | | Дальнейшее развитие подразумевает новую функциональность, такую как: добавление возможностидифференцировать меланому среди патологий, ускорив тем самым анализ спектра врачом, разработка графического интерфейса для удобства анализа возможных патологий, а также переход на более новые архитектуры классификаторов машинного обучения с увеличением точности определения патологий. | | |
| Апробация работы | | полученные результаты были апробированы на всероссийских и международных конференциях.   * выступление и публикация в сборнике международной конференции SPIE PhotonicsEurope 2020; * публикация в сборнике и диплом участника на ХLV Самарской областной научной студенческой конференции и публикации в сборнике конференции; * финалист конкурса молодых ученых «Умник» * участие и публикация в сборнике конференции "Информационные технологии в моделировании и управлении: подходы, методы, решения"; * участие на студенческом форуме «Инженерные кадры- будущее инновационной экономики России» ; * финалистконкурса «Росмолодежь 2019» ; * Участие в школе проектов Тольяттинского государственного университета «Лестница успеха»; * участие и публикация в сборнике научной конференции «Студенческие дни науки ТГУ». | | |
| План реализации проекта | | | | |
| Наименование и описание мероприятия (этапа) проекта | Сроки начала и окончания | | | Ожидаемые итоги |
| Анализ предоставленных спектров | 31.03.2020-01.05.2020 | | | Разработан ряд методов предобработки спектров КР и АФ |
| Анализ существующих алгоритмов для решения задачи классификации | 31.03.2020-01.05.2020 | | | Выбраныклассификаторы, показывающие наилучшие результаты в задачах дифференциации спектров |
| Анализ библиотек и фреймворков используемых в в реализации алгоритмов классификации машинного обучения | 01.05.2020-01.06.2020 | | | Выбран наиболее популярный и функциональный фреймворк для задач машинного обучения |
| Подготовка обучающей и тестовой выборки спектров тканей кожи | 01.06.2020-01.07.2020 | | | Программная реализация методовпредобработки |
| Построение ансамбля классификаторов машинного обучения | 01.07.2020-01.11.2020 | | | Разработана программная реализация ансамбля классификаторов |
| Тесты полученной модели и корректировка гиперпараметровалгоритмов машинного обучения | 01.11.2020-01.01.2021 | | | Конечный вариант модели |
| Создание интерфейса ПО | 01.01.2021-31.03.2021 | | | Разработанный программный модуль с пользовательским интерфейсом |

**СМЕТА**

**РАСХОДОВ НА РЕАЛИЗАЦИЮ ПРОЕКТА**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № п/п | Наименование статьи расходов | Единица - (чел., мес., шт. и т.п.) | Кол-во | Цена (руб.) | Стоимость |
| 1 | Виртуальная машина “Яндекс Облако” | мес. | 1 | 34612.40 | 34612.40 |
| 2 | Среда разработки PyCharm – годовая подписка | шт. | 1 | 15 632,45 | 15 632,45 |
| 3 | Разработка ПО | Чел./час | 1979 | 350 | 692650 |