Санкт–Петербургский государственный университет

КОВАЛЕНКО Лев Алексеевич

Выпускная квалификационная работа

Детектирование и классификация дефектов нефтепроводов на основе данных внутритрубных роботов-дефектоскопов

Уровень образования: бакалавриат
Направление 01.03.02 «Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2016 «Прикладная математика, фундаментальная информатика и программирование»
Профиль «Математическое и программное обеспечение вычислительных машин»

Научный руководитель:
доцент, кафедра технологии программирования,
к.т.н. Блеканов Иван Станиславович

Рецензент:
ООО ИТСК,
Кононов Ярослав Сергеевич

Санкт-Петербург
2020 г.

Содержание

Введение

В промышленности на сегодняшний день появилось огромное количество программных комплексов и информационных систем, позволяющих автоматизировать и оптимизировать производство. Такая тенденция подталкивает различные промышленные компании к внедрению инноваций в технологические процессы и проведению различных исследований.

Большой толчок в развитии получили многие области промышленности, например, автомобильная или нефтегазовая промышленность. Примером этому может служить то, что за последние несколько лет в группе компаний «Газпром Нефть» были внедрены инновационные решения для оптимизации процессов [1]:

- Инструмент для планирования очистки призабойной зоны для увеличения точности прогнозирования ожидаемого эффекта;
- Проект «Умная логистика» позволяет оптимизировать логистику поставки нефти и упростить контроль;
- Информационной система «ЭКСПРЕСС - ОЦЕНКА ГЕОЛОГИЧЕСКОГО СТРОЕНИЯ И ВЕРИФИКАЦИЯ ДАННЫХ», автоматизирующая процесс оценки геологического строения и верификации данных, построение карт, таблиц, схем геологических характеристик активов компании [2].

Кроме внедрения крупных проектов стоит отметить рост количества научно-исследовательских и опытно-конструкторских работ. С 2016 по 2018 их количество увеличилось в 24 раза. Исследования в компании «Газпом Нефть» проводятся по 14 различным направлениям [3].

Одним из важных направлений для автоматизации является оценка состояния нефтепроводов [4]. В группе компаний «Газпром Нефть» эксплуатируются примерно 12 тыс. км промысловых трубопроводов. В ходе эксплуатации нефтепроводы неизбежно изнашиваются. Это приводит к прорывам и утечкам нефти, что вызывает отрицательные экономические, экологические и репутационные эффекты. Во избежание этого проводятся работы по внутритрубной диагностике(ВТД) - съемке магнитограмм,

на основе которых эксперты могут оценить состояние нефтепровода. Следующим этапом по развитию этого направления является автоматизация процесса оценки состояния трубопровода с целью уменьшения затрат по временным ресурсам и увеличению качества процесса.

Практическая значимость заключается в том, что полученный в ходе работы программный комплекс может быть использован на предприятии для оптимизации процессов ВТД.

На данный момент промежуток времени от проведения съемки магнитограммы до получения отчета составляет до 1 месяца на 10 км трубы. Использование данного инструмента в его текущей реализации позволит частично автоматизировать процесс интерпретации магнитограмм. Это позволит сократить временной интервал от проведения съемки магнитограммы до получения отчета о состоянии трубопровода. При некоторой доработке и улучшении качества программного комплекса данный процесс может быть полностью автоматизирован.

Также стоит заметить, что за счет сокращения времени на расшифровку магнитограмм, появится возможность проводить ремонтные работы превентивно, что уменьшит количество прорывов нефтепроводов.

Цель работы

Цель данной работы - реализовать программный комплекс, позволяющий производить автоматическое детектирование и классификацию дефектов на основе анализа снимков магнитограмм.

Задачи работы

Для достижения цели были поставлены следующие задачи:

- Провести обзор существующих решений;
- Собрать и подготовить данные для дальнейших исследований: перевести магнитограммы из бинарного формата в табличный, провести точную разметку на основе представленных отчетов о состоянии нефтепроводов;
- Рассмотреть различные подходы и методы машинного обучения для детектирования швов и дефектов на различных магнитограммах;
- Разработать собственный подход анализа в виде модификации и комбинаций существующих алгоритмов;
- Выбрать стек технологий и спроектировать архитектуру программного комплекса;
- Разработать программное, решение позволяющее проводить оценку состояния трубопровода;
- Провести тестирование ПО и оценку качества разработанного подхода.

Глава 1. Обзор существующих решений и подходов в области анализа состояния инженерных сооружений

1.1 Обзор технологических решений

На сегодняшний день большое количество нефтегазовы сервисных компаний предлагает проведение диагностики состояния трубопровода с использованием магнитных и ультразвуковых устройств. Для проведения мониторинга состояния трубопроводов используются автономные зонды (Рис. 1.), оборудованные профилометрами и магнитными дефектоскопами. Такой зонд обычно имеет цилиндрическую форму, и по его бокам через равные угловые промежутки установлены измерительные блоки. Во время сервисных операций автономный зонд загружается в нефтепровод, а затем перемещается внутри вместе с нефтяным потоком, попутно замеряя магнитную индукцию вдоль стенок трубы. Полученные данные записываются в виде магнитограммы. Для ультразвуковой диагностики зонд оснащается другим типом датчиков, основанных на принципах ультразвукового замера толщины.

Рис. 1: Зонд для проведения внутритрубной диагностики

В сервисной компании «Baker Huges» предлагаются услуги по проведению магнитной и ультразвуковой внутритрубной дефектоскопии [5]. Для проведения диагностики используются магнитный и ультрозвуковой

дефектоскопы. Для интерпретации результатов дефектоскопии разработано ПО внутреннего пользования, позволяющее визуализировать магнитограмму и проводить ручную разметку.

Компания «Интрон плюс» также предлагает внутритрубную диагностику на основе данных магнитных дефектоскопов [6]. В компании разработано внутреннее программное обеспечение для визуализации и ручной разметки магнитограммы.

Одним из лучших решений на рынке является решение «Транснефти». У компании существует целый парк внутритрубных инспекционных приборов [7]. Магнитные дефектоскопы продемонстрированы на рисунке 2.

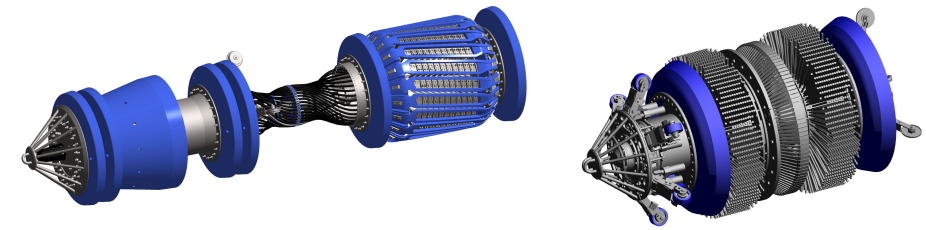


Рис. 2: Магнитные дефектоскопы

Стоит заметить, что ручной анализ магнитограммы занимает длительное время, так как размер одного участка трубопровода может составлять десятки километров. На анализ магнитограмм такого размера у экспертов уходит несколько месяцев.

Информации о существующих и эксплуатируемых программных решений для автоматизации процесса интерпретации и детектирования дефектов в открытом доступе не обнаружено.

1.2 Обзор алгоритмов анализа данных о состоянии сети трубопроводов

Для анализа состояния сети трубопроводов в статье [8] предлагается использовать статистические методы и методы машинного обучения. Сам анализ представляет собой детектирование дефектов и конструктивных элементов на магнитограмме, то есть классификацию участка трубы

как дефект, конструктивный элемент или полотно трубы. Для решения этой задачи стоит в первую очередь рассмотреть алгоритмы классификации:

- Random Forest Classifier – модель машинного обучения, основанная на применение деревьев решений. В основе этого ансамбля лежит подход бэкинга - обучение одинаковых моделей на разных подвыборках набора данных. Алгоритм использует усреднение для повышения точности прогнозирования и контроля соответствия [9];
- Extra Tree Classifier – данный алгоритм имеет такой же принцип, как и описанный выше Random Forest Classifier, но в качестве базовой модели использует рандомизированное решающее дерево [10];
- SVM Classifier – метод опорных векторов для классификации. Исходные векторы переводятся в пространство более высокой размерности, а затем в этом пространстве ищется разделяющая гиперплоскость с максимальным зазором, т.е. находящаяся максимально далеко от точек всех представленных классов [11];
- Logistic Regression - статистическая модель, позволяющая прогнозировать вероятность появление некоторого события по значениям множества признаков. Вероятность вычисляется путем сравнения события с логистической кривой [12].

В статьях [13], [14] описывается использование нейронных сетей для решения подобных задач. В них фигурируют сверточные нейронные сети и сети прямого распространения:

- Нейронные сети прямого распространения - многослой перцептрон, более подробно алгоритм описан в статье [15];
- Сверточные нейронные сети - алгоритм на основе применения механизма свертки для выделения признаков и дальнейшей классификации на их основе [16].

Основываясь на природе данных - сигналов магнитометров - было принято решение рассмотреть алгоритмы обработки сигналов, в частности алгоритмы фильтрации:

- Фильтр Калмана - эффективный рекурсивный фильтр, который проводит оценку вектора состояния динамической системы на основе ряда неполных и зашумленных данных [17];
- Вейвлет фильтр - фильтр, основанный на применении вейвлет - преобразования к набору сигналов [18].

1.3 Обзор методов классификации дефектов

Согласно ГОСТу [19] существуют следующие группы внутренних дефектов на полотне трубы:

- Трещины и зоны трещин, включая стресс-коррозионные трещины;
- Коррозия;
- Механические повреждения;
- Металлургические дефекты;
- Дефекты геометрии трубы.

Внутритрубная диагностика с помощью магнитных дефектоскопов позволяет выявлять коррозионные дефекты. Для этой группы дефектов существует способ классификации на основе физических параметров [20].

Таблица 1: Классы дефектов

Класс	Определение размеров
Общая коррозия	[W >= 3A] и [L >= 3A]
Питтинг	([1A <= W < 6A] и [1A <= L < 6A] и[0.5 < L/W < 2]) и не ([W >= 3A] и [L >= 3A])

Продольная канавка	[1A <= W < 3A] и [L/W >= 2]
Поперечная канавка	[L/W <= 0.5] и [1A <= L < 3A]
Продольная риска	[0 < W < 1A] и [L <= 1A]
Поперечная риска	[W <= 1A] и [0 < L < 1A]

L – длина дефекта, т.е. максимальный размер в продольном направлении (вдоль оси трубы); W – ширина дефекта, т.е. максимальный размер в поперечном (окружном) направлении; $A = 10$, если $\delta < 10$, $A = \delta$, если $\delta > 10$, где δ – номинальная толщина стенки трубы.

1.4 Обзор метрик качества алгоритмов по детектированию сварных швов и дефектов

Для оценки результатов были выбраны стандартные метрики: Recall, Precision и F1-measure. Эти метрики основываются на подсчете confusion matrix (матрицы ошибок), приведенной в таблицах 2 и 3.

Таблица 2: Матрица ошибок для сварных швов

	Сварной шов по разметке	Дефекты и структурные элементы по разметке
Предсказание класса сварного шва	True positive	False positive
Предсказание класса дефектов и структурных элементов	False negative	True negative

Таблица 3: Матрица ошибок для дефектов

	Дефекты по разметке	Сварные швы и структурные элементы по разметке
Предсказание класса дефекта	True positive	False positive
Предсказание класса сварных швов и структурных элементов	False negative	True negative

True positive – области, которые алгоритм определил как швы / дефекты, на самом деле являющиеся ими в размеченной выборке.

False positive – области, которые алгоритм определил как швы / дефекты, НЕ являющиеся ими в размеченной выборке.

False negative – области, которые алгоритм определил как НЕ швы / дефекты, на самом деле являющиеся ими в размеченной выборке.

True negative – области, которые алгоритм определил как НЕ швы / дефекты, НЕ являющиеся ими в размеченной выборке.

На основе представленной матрицы можно подсчитать выбранные метрики по следующим формулам:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Стоит отметить, что recall считается наиболее важной метрикой, поскольку большую ценность имеет выделение всех сварных швов и дефектов, которые присутствуют в экспертной разметке.

Глава 2. Разработка программного комплекса

2.1 Проектирование архитектуры решения и выбор стека технологий

Для реализации программного комплекса был выбран следующий стек:

- Python 3 - основной язык для реализации модели машинного обучения и веб сервиса [21];
- C++ - низкоуровневый язык для реализации парсера [22];
- React JS - фреймворк для реализации пользовательского интерфейса [23];
- Sqlite3 - база данных, используемая в проекте [24];
- Scikit-Learn - библиотека для реализации моделей машинного обучения [25];
- Pandas - библиотека для работы с табличными данными [26];
- SciPy - библиотека, реализующая методы оптимизации и фильтрации [27];
- Flask - фреймворк для реализации веб-сервиса [28];

Также для проекта была спроектирована архитектура. Ее схема продемонстрирована на рисунке 3. Архитектура системы разделена на 5 компонентов: интерфейс пользователя, API, загрузчик данных, модель машинного обучения и база данных. Данная архитектура поддерживает горизонтальное масштабирование приложения. Более подробное описание архитектуры представленно ниже:

- Интерфейс пользователя позволяет визуализировать данные магнитограммы, проводить ручную и автоматическую разметку, оценку и коррекцию автоматической разметки. Логически его можно разбить

на 5 частей: компоненту визуализации данных, таблицы для швов и дефектов, интерфейсы загрузчика данных и взаимодействия с моделью. Компонента визуализации магнитограммы позволяет демонстрировать участок магнитограммы с помощью двухмерной heatmap. Таблицы для швов и дефектов отображают результаты разметки и позволяют корректировать ее результаты, а также дают возможность демонстрировать конкретные швы или дефекты. Интерфейс загрузчика данных реализует возможность загружать данные магнитограмм в бинарном формате. Интерфейс модели необходим для задания гиперпараметров фильтрации и запуска моделей.

- Компонента API необходима для связи пользовательского интерфейса со всей остальной системой. Модуль реализован в виде restful вебсервиса с помощью фреймворка Flask. Внутри API находятся три интерфейса для работы с базой данных, моделью и загрузчиком данных.

- Загрузчик данных занимается расшифровкой бинарных файлов магнитограмм в соответствии с заданным форматом, преобразованием данных из формата WideData в TidyData формат и загрузке данных в базу данных. Загрузчик был реализован на языке C++, так как была необходимость в расшифровке бинарных файлов и непосредственной работой с битами.

- Модуль модели необходим для автоматической разметки магнитограммы по швам и дефектам с дальнейшей загрузкой их в базу данных. Для реализации модели машинного обучения применялся язык Python 3 и набор библиотек: Pandas, Scikit-Learn и SciPy.

- База данных - система для хранения данных программного комплекса в табличном виде. В реализации прототипа использовалась sql база данных Sqlite3.

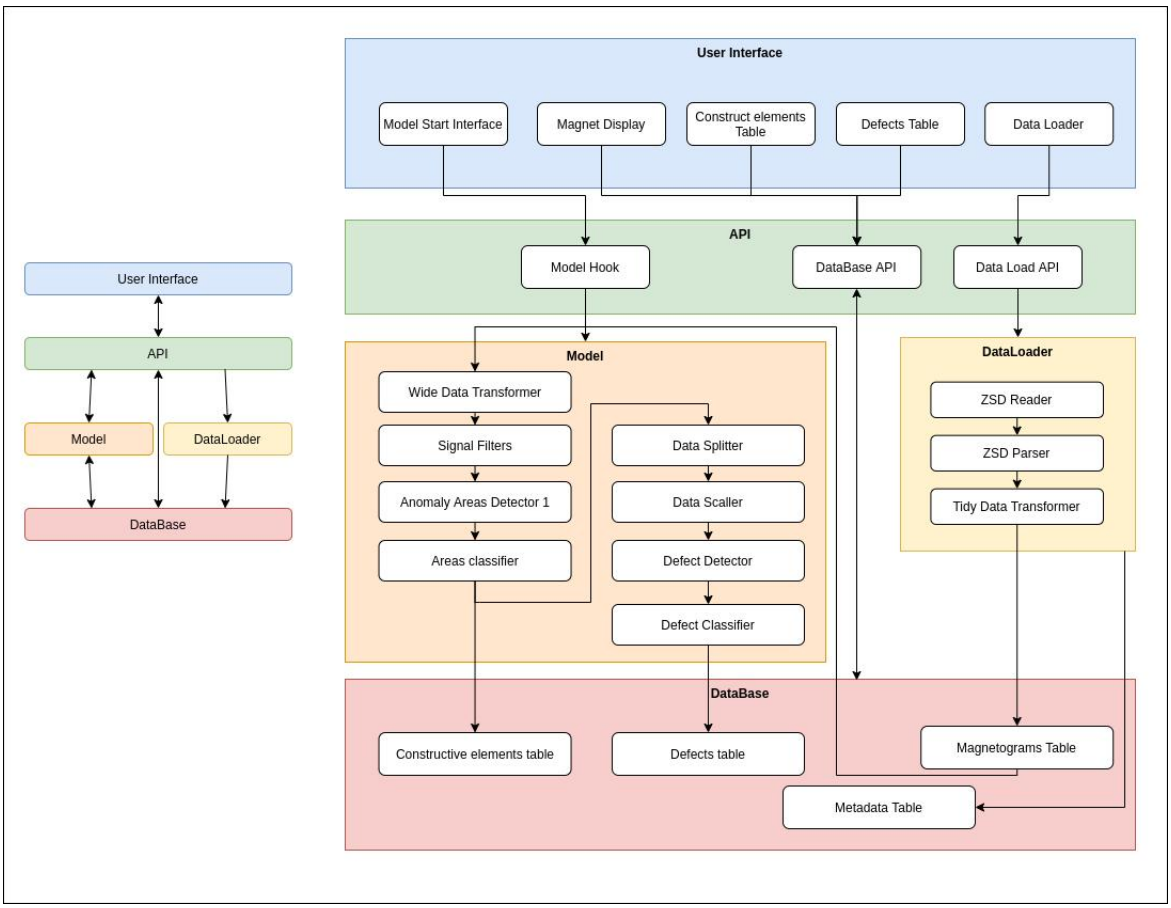


Рис. 3: Архитектура программного комплекса

2.2 Структура и особенности данных

Исходные данные зондов находились в бинарном формате. Для анализа и построения математической модели эти данные были расшифрованы и преобразованы в табличный формат «.csv» с помощью специально разработанного скрипта.

Для каждой магнитограммы был следующий набор файлов: исходный файл магнитограммы в бинарном виде, преобразованный табличный файл магнитограммы «.csv» и экспертный отчет «.pdf» или «.doc». В свою очередь, в отчете находилась информация об условиях съемки, а также давалась экспертная разметка магнитограммы с указанием конструктивных элементов и дефектов, включая изображения соответствующих примеров. Всего в изначальном датасете было 36 магнитограмм, из которых 33 оказались уникальными (3 оставшиеся соответствовали повторному обследо-

ванию одних и тех же труб и не содержали дополнительной информации). Список параметров, полученных после расшифровки магнитограмм, приведен в Таблице 4.

Таблица 4: Параметры магнитограмм

Параметр	Описание	Единица измерения
Tag	Начало новой записи данных (технический параметр)	-
Size	Размер поля для записи данных (технический параметр)	-
Time	Момент записи	мкс
Dist	Значение дистанции	10 мкм
Index	Порядковый номер измерения	-
Flags	Флаги сканирования (технический параметр)	-
Status X	Статус для блока из четырех датчиков, описывающий возможные ошибки в них: 0 – отсутствие ошибок; X – номер блока	-
X.Y	Значение датчика Y из блока X, например, «1.3»	-

Термин «Технический параметр» означает, что данная переменная используется либо для расшифровки, либо для конверсии из исходного бинарного файла. Из приведенного списка для анализа магнитограмм наиболее важными факторами являются Dist и X.Y, так как именно они указывают, на каком расстоянии находится дефект/шов/структурный элемент и каким измеренным значениям он соответствует.

Первичный анализ данных показал необходимость перепроверки и переразметки данных. Параметры, использованные для разметки сварных швов, дефектов и структурных элементов, приведены в Таблицах 5 и 6.

Таблица 5: Параметры конструктивных элементов

Параметр	Описание	Единица измерения
id	Уникальный идентификатор конструктивного элемента – строка	-
contructive type	Тип конструктивного элемента на основе отчетов	-
start dist	Начало конструктивного элемента	10 мкм
end dist	Конец конструктивного элемента	10 мкм
wall width	Толщина стенки трубы на основе отчета	мм

Таблица 6: Параметры дефектов

Параметр	Описание	Единица измерения
id	Уникальный идентификатор дефекта – строка	-
defect type	Тип дефекта на основе отчетов	-
defect place	Местоположение дефекта: «Стенка трубы» или «Сварной шов»	-
defect depth	Глубина дефекта на основе отчета	%
start dist	Начало дефекта	10 мкм
end dist	Конец дефекта	10 мкм

После переразметки в соответствующие папки магнитограмм были добавлены файлы конструктивных элементов *construct_element.csv* и файл разметки дефектов *defect.csv*.

При сопоставлении исходных данных магнитограмм с экспертной разметкой, приведенной в отчетах, было обнаружено расхождение длины пути зонда на расстояние вплоть до 150 см, что, в свою очередь, привело к отли-

чию координат дефектов и конструктивных элементов между датасетом и сопутствующим отчетом. Предположительно, причиной этого расхождения является ошибка в ПО, используемом для генерации графиков к отчету. Для экспертного отчета подобное расхождение является нежелательным, но не критичным, так как ремонт поврежденного участка происходит не точечно, а в рамках целой секции трубы. В то же время такое несоответствие между отчетом и магнитограммой весьма негативно сказывается на точности математической модели, которая не может обучиться на противоречащих данных. По этой причине силами сотрудников Дирекции Инновационного Развития была выполнена ручная проверка и переразметка датасета каждой магнитограммы («.csv»), а обновленные файлы были добавлены в соответствующие папки.

Важно заметить, что число записей (строк с показаниями датчиков) в магнитограмме зависит от пройденного зондом расстояния, которое может достигать 40 км. В представленном ниже датасете длина составляла 1.5-10 км. Участки свыше 10 км были исключены из рассмотрения в связи с трудоемкостью уточнения изначальной разметки. В результате итоговый датасет составил 22 магнитограммы для анализа швов и 18 для анализа дефектов (некоторые трубы не имели дефектов согласно отчетам). Итоговый список магнитограмм для построения математической модели приведен в Таблице 7.

Таблица 7: Итоговый список магнитограмм

Магнитограмма	Количество записей
Магнитограмма №1	2125866
Магнитограмма №2	1470101
Магнитограмма №3	3815418
Магнитограмма №4	2482089
Магнитограмма №5	3761227
Магнитограмма №6	1072643
Магнитограмма №7	449613
Магнитограмма №8	711720

Магнитограмма №9	412681
Магнитограмма №10	1165283
Магнитограмма №11	3047114
Магнитограмма №12	575441
Магнитограмма №13	1498927
Магнитограмма №14	683275
Магнитограмма №15	1108527
Магнитограмма №16	1630936
Магнитограмма №17	338127
Магнитограмма №18	233018
Магнитограмма №19	291116
Магнитограмма №20	222167
Магнитограмма №21	860107
Магнитограмма №22	1772941

Для выявления особенностей данных были визуализированы сигналы для всех магнитограмм. Поскольку данные писались одновременно 64 независимыми датчиками, то для отображения графиков использовалось усредненное значение по всем датчикам, записанное на данной координате (для сварных швов и структурных объектов). Для дефектов строились индивидуальные графики для каждого датчика. В итоге были зафиксированы следующие результаты: рисунок 4, рисунок 5 и рисунок 6.

Определены паттерны конструктивного элемента «Сварной шов» и их средние показатели (ширина 5 сантиметров и высота 250 у.е.) (Рис. 4).

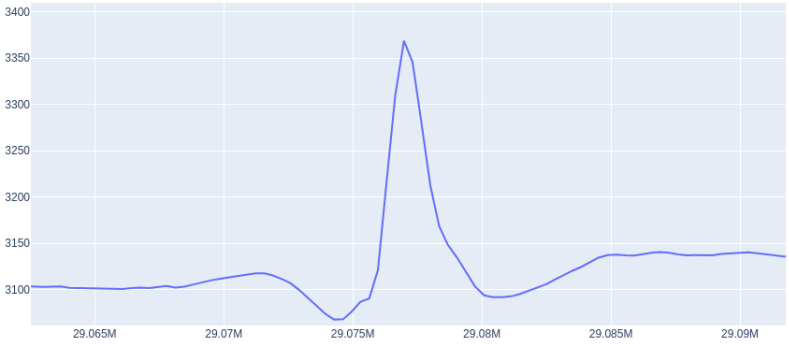
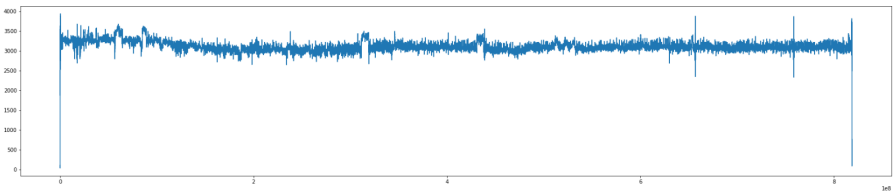
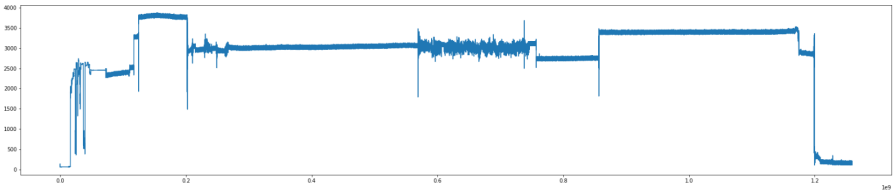


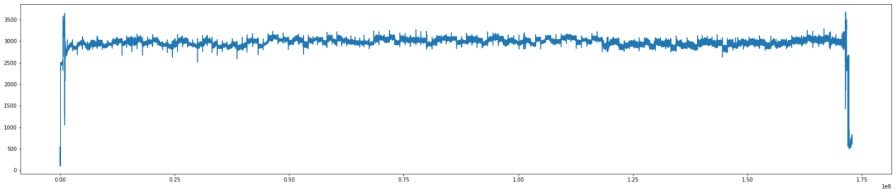
Рис. 4: Выделенный паттерн шва



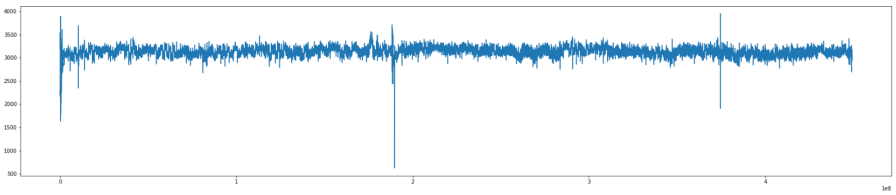
Магнитограмма №1



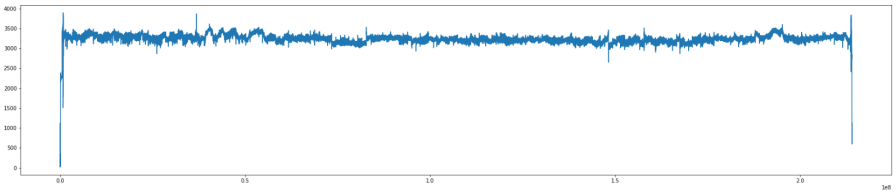
Магнитограмма №3



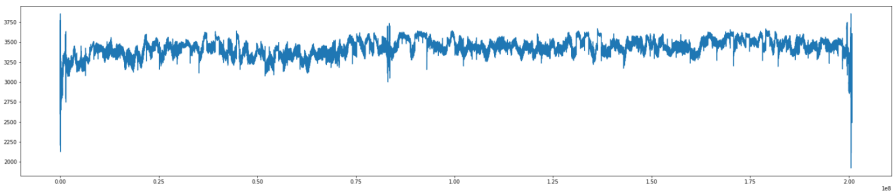
Магнитограмма №9



Магнитограмма №10

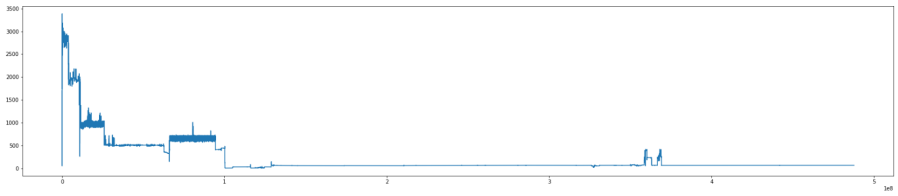


Магнитограмма №12

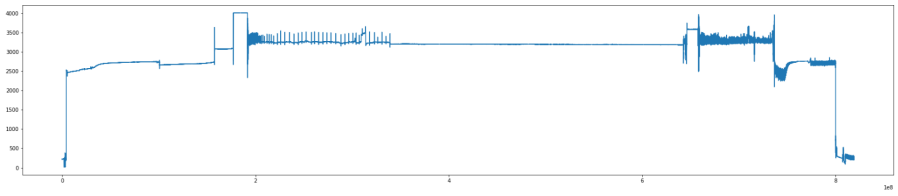


Магнитограмма №14

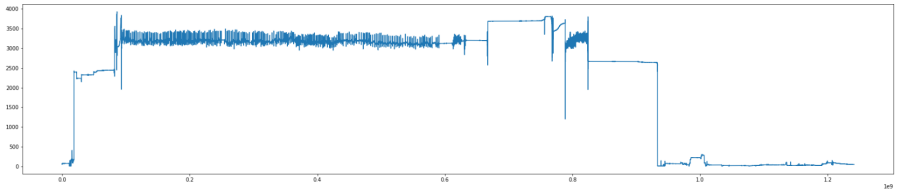
Рис. 5: Магнитограммы с сильными шумами



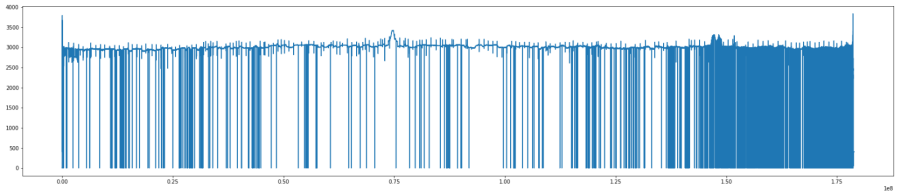
Магнитограмма №2



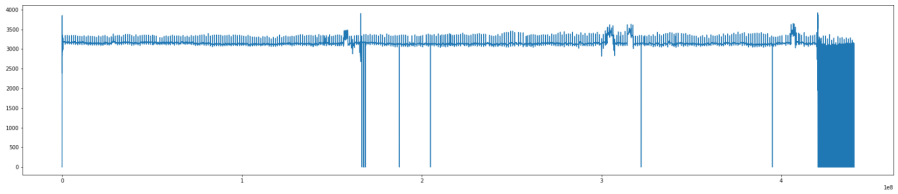
Магнитограмма №4



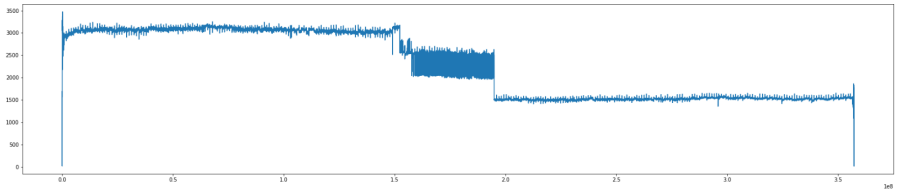
Магнитограмма №5



Магнитограмма №7



Магнитограмма №15



Магнитограмма №16

Рис. 6: Магнитограммы с битыми или недостоверными данными

Зашумленные и «битые» магнитограммы также использовались для

построения математической модели, но с заведомым ожиданием ухудшения качества детектирования на них.

2.3 Процесс предобработки данных

При анализе статусов блоков датчиков было замечено, что почти для половины магнитограмм датчики выдавали сообщения об ошибках. Поскольку эксперты все же смогли разметить подобные «битые» данные, было принято решение не исключать проблемные магнитограммы и попробовать восстановить их двумя разными способами:

- Анализ показаний статусов блоков датчиков. Данный подход оказался неудачным, ввиду того, что для некоторых магнитограмм количество ненулевых статусов соответствовало количеству выполненных измерений. Соответственно, во всей магнитограмме были отражены недостоверные данные.

- Проведение порогового анализа. На основе магнитограмм, визуально не имеющих битых данных, была рассчитана нижняя граница доверительного интервала показания датчиков. Определение «битых» / недостоверных данных проводилось с помощью рассчитанного порога: в среднем значения в магнитограмме по оси ординат колебались от 0 до 4000 у.е. (условных единиц); значения ниже 1000 у.е. экспертно были признаны недостоверными и заменены на арифметическое среднее по достоверным данным этой магнитограммы. Наглядные примеры различных «битых» и зашумленных магнитограмм продемонстрированы на рисунках 5 и 6.

Таким образом, метод порогового анализа позволил использовать для работы весь датасет, включая магнитограммы, состоящие только из «битых» данных.

2.4 Построение математической модели

Для решения задач был выбран следующий подход:

- Разработать аналитическое решение, способное определять области, схожие с паттерном сварного шва;
- Уточнить решение с помощью модели машинного обучения, обучив ее на подготовленной выборке;
- Провести разделение магнитограммы на секции по детектированным швам;
- Для каждой выделенной секции с помощью аналитического решения найти области, являющиеся локальными минимумами значений сигналов;
- Уточнить решение с помощью фильтрации и модели машинного обучения, обучив ее на подготовленной выборке.

2.5 Выбор признаков

Для обучения модели по детектированию швов на сформированном датасете в качестве признака было выбрано окно размером в 17 показаний на усредненном по всем датчикам сигнале: 8 ближайших показаний левее заданной точки, показание на данной точке, 8 показаний правее. Размер окна был выбран эмпирическим образом после проведения экспериментов для максимизации значений метрик recall и precision.

Для обучения модели по детектированию дефектов на сформированном датасете в качестве признака было выбрано окно размером в 20 показаний на всех датчиках: 9 ближайших показаний левее заданной точки, показание для данной точки, 10 показаний правее. В результате выбора такого окна была получена матрица признаков размером $n \times 20$, где n – количество датчиков. Для увеличения количества сэмплов с дефектами, производился циклический сдвиг полученной матрицы построчно: изначально строка с дефектом размещалась «внизу» матрицы, а затем на каждой итерации смещалась «вверх», пока не оказывалась на первом месте. Подобная манипуляция была нужна, чтобы проиллюстрировать принципиальную возможность нахождения дефекта в любом угловом секторе

трубы. В противном случае, в процессе обучения модель могла бы решить, что дефекты всегда располагаются исключительно в определенном секторе, а похожие показания в других секторах – статистические выбросы. После подготовки всех семплов полученные матрицы подверглась преобразованию в массив путем последовательной записи столбцов. Аналогично сварным швам размер окна был выбран эмпирически, исходя из экспериментов для максимизации значения метрик recall и precision у модели машинного обучения.

2.6 Решение по детектированию швов

Аналитическое решение представляет собой алгоритм по поиску характерных пиков, явно выбивающихся из основного сигнала, и, согласно картам разметки, соответствующих сварным швам. Основной проблемой такого подхода является наличие битых данных и сильная зашумленность некоторых магнитограмм. Для решения проблемы зашумленных данных были предприняты следующие шаги:

- Использование среднего значения показаний датчиков для анализа;
- Подсчет скользящего среднего с окном размером в 100 показаний и вычет его из сигнала;
- Подсчет скользящего среднего с окном размером в 2 показания и вычет его из сигнала.

Также в ходе решения проблемы было рассмотрено применение фильтра Калмана и фильтра на основе вейвлет-преобразования. В финальную версию решения они не вошли, поскольку их применение к магнитограмме занимало значительно больше времени, чем весь остальной алгоритм аналитического решения. В частности, для одной магнитограммы длиной 1 км время работы фильтров составило:

- Вейвлет фильтр – 123 минуты;
- Фильтр Калмана – 36 минут.

После применения описанных преобразований к усредненному сигналу производился поиск пиков с помощью функции find peaks, реализованной в библиотеке SciPy. Для отсечения шумов использовались пороговые значения по относительной и абсолютной высоте пика. Для работы было определено пороговое значение, равное 1000.

Важно заметить, что данные фильтры использовались исключительно для аналитического решения, чтобы эффективнее выделять пики и находить их максимальные значения. При этом профиль самих пиков искажался, что делало невозможным их дальнейшее использование для обучения математической модели (разрабатываемая модель машинного обучения по-прежнему обучалась на зашумленных данных, впрочем, теперь с уже точно заданными координатами пиков).

Далее, после выделения областей интереса происходила фильтрация с помощью модели машинного обучения. Обучение модели проводилось на выделенных ранее признаках, соответсвующих двум классам: швам и полотну трубы. Баланс обучающей выборки был 1:1.

2.7 Решение по детектированию дефектов

Для аналитического решения был выбран следующий подход:

- Магнитограмма разбивалась на секции по швам;
- Для каждой секции производился поиск локальных минимумов значений сигналов датчиков;
- Производилась фильтрация выбранных минимумов по глубине.

Для поиска локальных минимумов производилась предобработка данных. Сначала данные интерполировались по дистанции для достижения сетки с равномерным шагом (т.е. пространственно измерения были равноудалены друг от друга). Также применялся гауссовский фильтр с целью уменьшения зашумленности. Минимумы искались с помощью функции find peaks, реализованной в SciPy. Данная функция применялась к

среднему сигналу с каждого блока датчиков, умноженному на -1 для отражения сигнала относительно оси абсцисс (Ox) на рисунках 5 и 6. Также производилась фильтрация пиков по их метапараметрам – ширине и относительной высоте. Для каждого пика сохранялись данные об абсолютной и относительной высоте, ширине, координатах левой и правой границ пика. Также производился подсчет процента глубины дефекта от общего сигнала. Наконец, для фильтрации дефектов применялся пороговый фильтр по проценту глубины потери металла.

После выбора зон возможных дефектов производилось уточнение с помощью модели машинного обучения. Ее обучение проводилось подобным образом, как описано в пункте выше.

2.8 Решение по классифкации дефектов

Данное решение основывается на методических указаниях, данных в [19]. После детектирования дфектов проводился подсчет их физических параметров: длины и ширины. На основе данных параметров выбирался один из классов дефекта. Также был добавлен класс «Аномалия», соответсвующий случаю, когда ширина или длина дефекта равнялась нулю по результатам замеров.

Глава 3. Проведение эксперимента

3.1 Постановка эксперимента

Для валидации алгоритмов по детектированию швов и дефектов было поставлено 7 экспериментов:

- Валидация аналитического решения по детектированию швов;
- Валидация бинарного классификатора для швов на подготовленном датасете;
- Валидация алгоритма по детектированию швов;
- Валидация аналитического решения по детектированию дефектов;
- Валидация бинарного классификатора для швов на подготовленном датасете;
- Валидация алгоритма по детектированию швов;
- Тестирование алгоритмов на тестовой магнитограмме.

Для валидации аналитических решений использовался весь набор магнитограмм. На его основе проводилась кросс-валидация - цикл оценки решения. На каждой итерации цикла выбиралась одна магнитограмма для разметки, а на основе оставшихся производилось обучение модели. После получения разметки она сопоставляется с разметкой экспертов, рассчитывались метрики. Аналогичный подход использовался для валидации полных решений по детектированию швов и дефектов. Проверка бинарных классификаторов проводилась на основе заранее составленных на основе экспертной разметки обучающих выборок для каждой из магнитограмм. Проверка также представляла собой процесс кросс-валидации. На каждой итерации выбиралась одна из выборок в качестве тестовой, а на остальных проводилось обучение алгоритма. Далее проводилась оценка модели на тестовой магнитограмме по заданным метрикам.

Помимо данных, используемых для исследований, была выделена отдельная тестовая магнитограмма без разметки, которую требовалось интерпретировать, то есть предоставить информацию о расположении швов и дефектов. В дальнейшем полученная разметка сопоставлялась с трубными журналами.

3.2 Результаты экспериментов

Ниже представлены результаты проведенных экспериментов в виде таблиц с результатами кросс-валидации.

Таблица 8: Метрики для аналитического решения по детектированию швов

Магнитограмма	recall	precision	f1 score
Магнитограмма №1	0.914	0.297	0.448
Магнитограмма №2	0.889	0.002	0.004
Магнитограмма №3	0.821	0.013	0.026
Магнитограмма №4	0.951	0.355	0.517
Магнитограмма №5	0.986	0.636	0.773
Магнитограмма №6	0.953	0.677	0.792
Магнитограмма №7	0.939	0.173	0.292
Магнитограмма №8	0.963	0.225	0.365
Магнитограмма №9	0.951	0.172	0.291
Магнитограмма №10	0.813	0.12	0.209
Магнитограмма №11	0.955	0.717	0.819
Магнитограмма №12	0.738	0.101	0.178
Магнитограмма №13	0.929	0.315	0.47
Магнитограмма №14	0.714	0.164	0.267
Магнитограмма №15	0.972	0.433	0.599
Магнитограмма №16	0.979	0.774	0.865
Магнитограмма №17	0.945	0.448	0.608
Магнитограмма №18	0.885	0.327	0.478
Магнитограмма №19	0.988	0.756	0.857
Магнитограмма №20	0.934	0.337	0.495

Магнитограмма №21	0.962	0.852	0.904
Магнитограмма №22	0.964	0.705	0.814
Среднее	0.91568	0.39086	0.54786

Как видно из расчетов в таблице 8, аналитическая модель обладает высоким параметром recall (0.714-0.988, среднее 0.91568), но также достаточно низким precision (0.002-0.705, среднее 0.39086). Такой разброс precision продемонстрировал необходимость уточнения решения.

Таблица 9: Кросс-валидация на Random Forest Classifier на швах

Магнитограмма	recall	precision	f1 score
Магнитограмма №1	0.876747	0.949106	0.911493
Магнитограмма №2	0.444444	0.8	0.571428
Магнитограмма №3	0.934959	0.787671	0.855018
Магнитограмма №4	0.896341	0.97351	0.933333
Магнитограмма №5	0.969388	0.996503	0.982759
Магнитограмма №6	0.994083	0.973913	0.983895
Магнитограмма №7	0.807107	0.969512	0.880887
Магнитограмма №8	0.950617	0.888462	0.918489
Магнитограмма №9	0.993827	0.899441	0.944281
Магнитограмма №10	0.972906	0.833333	0.897727
Магнитограмма №11	0.988839	0.977925	0.983352
Магнитограмма №12	0.907692	0.907692	0.907692
Магнитограмма №13	0.855457	0.870871	0.863095
Магнитограмма №14	0.980583	0.971154	0.975846
Магнитограмма №15	0.997245	1	0.998621
Магнитограмма №16	1	1	1
Магнитограмма №17	0.912088	0.954023	0.932584
Магнитограмма №18	0.948718	0.986667	0.96732
Магнитограмма №19	1	1	1

Магнитограмма №20	0.934211	0.898734	0.916129
Магнитограмма №21	0.981191	0.96904	0.975078
Магнитограмма №22	0.978088	0.983968	0.981019
Среднее	0.923842	0.935978	0.92987

По результатам эксперимента можно сделать заключение о применимости Random Forest Classifier в качестве оптимизатора, так как минимальное значение precision составляет 0.78 для модели. К тому же сильное падение метрик фиксируется только на магнитограммах, отнесенных к зашумленным магнитограммам или магнитограммам с «битыми» данными.

Таблица 10: Метрики для итогового решения по детектированию швов

Магнитограмма	recall	precision	f1 score
Магнитограмма №1	0.85	0.69	0.76
Магнитограмма №2	0.71	0.31	0.43
Магнитограмма №3	0.67	0.70	0.68
Магнитограмма №4	0.84	0.71	0.77
Магнитограмма №5	0.89	0.99	0.94
Магнитограмма №6	0.90	0.98	0.94
Магнитограмма №7	0.80	0.99	0.88
Магнитограмма №8	0.92	0.54	0.68
Магнитограмма №9	0.93	0.44	0.6
Магнитограмма №10	0.81	0.26	0.39
Магнитограмма №11	0.96	0.82	0.88
Магнитограмма №12	0.73	0.57	0.64
Магнитограмма №13	0.74	0.72	0.73
Магнитограмма №14	0.64	0.93	0.76
Магнитограмма №15	0.97	0.86	0.91
Магнитограмма №16	0.96	0.97	0.96
Магнитограмма №17	0.86	0.80	0.83

Магнитограмма №18	0.86	0.76	0.81
Магнитограмма №19	0.99	0.94	0.96
Магнитограмма №20	0.90	0.96	0.93
Магнитограмма №21	0.96	0.98	0.97
Магнитограмма №22	0.96	0.97	0.96
Среднее	0.856	0.767	0.809

В процессе оценки гипотеза об ухудшении качества за счет добавления магнитограмм, отнесенных к магнитограммам с зашумленными или «битыми» данными, подтвердилась. Напротив, при рассмотрении магнитограмм, не имеющих аномалий в данных, можно говорить о довольно высоких показателях recall и precision (больше 75%).

Таблица 11: Результаты аналитического решения по детектированию дефектов

Магнитограмма	Найдено дефектов	Дефектов в разметке	Совпадений с разметкой
Магнитограмма №5	133	16	2
Магнитограмма №6	23824	17	15
Магнитограмма №7	17665	11	9
Магнитограмма №8	5847	9	3
Магнитограмма №9	852	130	15
Магнитограмма №10	20437	3	3
Магнитограмма №11	56637	80	68
Магнитограмма №12	2664	14	2
Магнитограмма №13	62	31	2
Магнитограмма №13	16726	49	13
Магнитограмма №14	12581	21	18
Магнитограмма №15	30631	15	14
Магнитограмма №17	3163	22	15
Магнитограмма №18	339	5	1

Магнитограмма №19	11283	21	17
Магнитограмма №20	1502	9	3
Магнитограмма №21	24151	24	19
Магнитограмма №22	56332	216	173
Сумма	284829	693	392

Стоит отметить, что оценка данного решения по метрикам затруднена за счет того, что экспертная разметка не является полной (при детальном анализе часть дефектов оказалась пропущенной экспертами при их разметке).

Таблица 12: Результаты аналитического решения по детектированию дефектов

Магнитограмма	recall	precision	f1 score
Магнитограмма №5	0.761719	0.884354	0.818468
Магнитограмма №6	0.244485	0.820988	0.37677
Магнитограмма №7	0.960227	0.522815	0.677015
Магнитограмма №8	0.946181	0.582265	0.7209
Магнитограмма №9	0.964063	0.526036	0.680669
Магнитограмма №10	1	0.75	0.857143
Магнитограмма №11	0.551172	0.868041	0.674233
Магнитограмма №12	0.975446	0.556688	0.70884
Магнитограмма №13	0.669962	0.601833	0.634073
Магнитограмма №14	0.924107	0.671351	0.777708
Магнитограмма №15	0.15	0.993103	0.260633
Магнитограмма №16	0.143145	0.662005	0.235391
Магнитограмма №17	0.892045	0.842953	0.866804
Магнитограмма №18	0.43125	0.901961	0.58351
Магнитограмма №19	0.320685	0.814745	0.460225
Магнитограмма №20	0.883681	0.538055	0.668857
Магнитограмма №21	0.427734	0.948052	0.589502

Магнитограмма №22	0.630859	0.913959	0.746469
Среднее	0.65982	0.74440	0.699563

Данный эксперимент показывает, что на хороших магнитограммах результаты Random Forest Classifier высоки и применимы для уточнения аналитической модели. Но в тоже время, на магнитогрммах низкого качества recall падает до 0.15.

Таблица 13: Результаты аналитического решения по детектированию дефектов

Магнитограмма	Найдено дефектов	Дефектов в разметке	Совпадений с разметкой
Магнитограмма №5	76	16	2
Магнитограмма №6	2123	17	15
Магнитограмма №7	2749	11	9
Магнитограмма №8	484	9	5
Магнитограмма №9	265	130	54
Магнитограмма №10	4206	3	1
Магнитограмма №11	2992	80	56
Магнитограмма №12	216	14	3
Магнитограмма №13	45	31	0
Магнитограмма №13	886	49	10
Магнитограмма №14	1412	21	18
Магнитограмма №15	215	15	1
Магнитограмма №17	84	22	5
Магнитограмма №18	61	5	3
Магнитограмма №19	284	21	15
Магнитограмма №20	102	9	1
Магнитограмма №21	73	24	2
Магнитограмма №22	483	216	56
Сумма	16756	693	256

В результате применения модели машинного обучения количество верно детектируемых дефектов сильно упало и составляет только третью часть замеченных дефектов, но при этом общее число детектируемых областей уменьшилось на порядок.

Наличие тестовой магнитограммы гарантировало невозможность переобучения модели на тестовой выборке и позволяло провести независимую верификацию качества разрабатываемой математической модели. Данная магнитограмма была успешно размечена алгоритмом и была выслана Заказчику для экспертной проверки. Всего на ней было обнаружено 481 сварных шва и 14 дефектов.

3.3 Выводы

По результатам экспериментов можно сказать о высоком качестве работы решения по детектированию швов как на магнитограммах хорошего качества, так и на магнитограммах, содержащих битые, недостоверные или зашумленные данные.

Решение по детектированию дефектов хорошо себя показывает на магнитограммах высокого качества, имеющих малое количество шумов. Некорректная идентификация большого числа дефектов обусловлена тем, что на некоторых магнитограммах присутствуют многочисленные шумы, которые, как и дефекты, являются локальными минимумами и определяются алгоритмом как дефекты.

Заключение

В рамках проведенных работ были выполнены следующие этапы:

- Проведен обзор существующих решений;
- Разработано программное обеспечение для преобразования файлов формата «.zsd» к табличному формату «.csv»;
- Собраны и подготовины данные для дальнейших исследований: магнитограммы перведены из бинарного формата в табличный, проведена точная разметка на основе представленных отчетов о состоянии нефтепроводов для 22 магнитограмм;
- Рассмотрены различные подходы и методы машинного обучения для детектирования швов и дефектов;
- Разработаны две модели для детектирования сварных швов и дефектов, результаты качественной оценки решений зафиксированы в разделе 3.2;
- Выбран стек технологий для реализации программного комплекса;
- Спроектирована архитектура программного комплекса;
- Реализован прототип цифрового продукта, производящий автоматическую разметку магнитограммы с дальнейшей возможностью ее корректировки в ручном режиме.

При дальнейшем развитии проекта и доработке решения следует обогатить выборку за счет добавления новых магнитограмм с экспертной разметкой. Также для улучшения работы самого решения существуют следующие возможности:

- Применить фильтрацию к данным;
- Провести более подробный анализ паттерна дефекта;

- Обсудить с экспертами выявленные аномальные ситуации (в том числе их возможную физическую природу);
- Исследовать применимость дополнительных data science подходов (бустинговые деревья в текущей постановке задачи, нейросети для обработки изображений и выявления на них паттернов швов/дефектов);
- Провести детальный подбор гиперпараметров для моделей;
- Добавить такие метапризнаки, как ширина, высота, глубина и т.п. для детектируемого объекта.

Список литературы

[1] Алексеев, А. Точки инновационного роста /А. Алексеев // «Сибирская нефть». – 2019. – №161. – С. 46-52.

[2] Официальный сайт ООО «ИТСК»: сайт. – URL: http://it-sk.ru/products/ (дата обращения: 06.03.2020). – Текст: электронный.

[3] Никоноров, А. Двигатель инноваций / А.Никоноров // «Сибирская нефть». – 2018. – №152. – С. 52-58.

[4] Шалай, В.В. Анализ технического состояния объектов линейной част магистральных Нефтепроводов, определение оптимальных способов поддержания объектов линейной части в нормативном состоянии / В.В. Шалай, М.М. Васильев, К.А. Шумаков // «Омский научный вестник»– Омск, 2004. – С. 196-199.

[5] Официальный сайт «Baker Huges»: сайт. – URL: https://www.bakerhughes.com/ (дата обращения: 06.03.2020). – Текст: электронный.

[6] Официальный сайт «Интрон gk.c»: сайт. – URL: https://www.intron.ru/ (дата обращения: 06.03.2020). – Текст: электронный.

[7] Парк приборов для провдения ВТД компании «Транснефть»: сайт. – URL: https://diascan.transneft.ru/klientam/vnytritrybnaya-diagnostika/park-vnytritrybnih-inspekcionnih-priborov/?preview=1 (дата обращения: 06.03.2020). – Текст: электронный.

[8] Закирзаков А.Г., Егоров А.Л. АНАЛИЗ СОСТОЯНИЯ СЕТИ МАГИСТРАЛЬНЫХ НЕФТЕПРОВОДОВ ТЮМЕНСКОЙ ОБЛАСТИ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ДАННЫХ // Современные проблемы науки и образования. – 2015. – №1-1. – URL: http://www.science-education.ru/ru/article/view?id=18926 (дата обращения: 23.04.2020).

[9] Чистяков С.П. СЛУЧАЙНЫЕ ЛЕСА: ОБЗОР // Труды Карельского научного центра РАН – 2013. – №1. – С. 117-136.

[10] Geurts, P., Ernst, D., Wehenkel, L. Extremely randomizedtrees // Machine Learning. – 2006. – №63. – p.3–42.

[11] Srivastava, D., L. Bhambhu Data classification using support vector machine. // Journal of Theoretical and Applied Information Technology. – 2010. – №12(1). – p. 1-7.

[12] Chao-Ying J. P., Lee K. L., Ingersoll G. M. An Introduction to Logistic Regression Analysis and Reporting. // The Journal of Educational Research. – 2002. – №96(1). – p. 3-14.

[13] Cheng J., Wang M. Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques // Automation in Construction. – 2018. – №95. – p. 155-171.

[14] Велькер Н.Н., Бондаренко А.В., Вершинин А.С., Дашевский Ю.А. ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ОБРАТНОЙ ЗАДАЧИ ВНУТРИТРУБНОЙ МАГНИТНОЙ ДЕФЕКТОСКОПИИ МАГИСТРАЛЬНЫХ ТРУБОПРОВОДОВ // Интерэкспо Гео-Сибирь. – 2018. – URL: https://cyberleninka.ru/article/n/primenenie-neyronnyh-setey-dlya-resheniya-obratnoy-zadachi-vnutritrubnoy-magnitnoy-defektoskopii-magistralnyh-truboprovodov (дата обращения: 23.04.2020).

[15] Murtagh F. Multilayer perceptrons for classification and regression // Neurocomputing. – 1991. – №2. – p.183-197.

[16] Indolia S., Goswami A. K., Mishra S.P., Asopa P. Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach // Procedia Computer Science. – 2018. – №132. – p.679-688.

[17] Дегтярев, А.А. Элементы теории адаптивного расширенного фильтра Калмана / А. А. Дегтярев, Ш. Тайль – Препринты ИПМ им. М. В. Келдыша, 2003 – 36 с.

[18] Демаков Н.В., Кузовников А.В., Пашков А.Е., Анжина В.А. ФИЛЬТРАЦИЯ СИГНАЛОВ С ПОМОЩЬЮ ВЕЙВЛЕТ-

ПРЕОБРАЗОВАНИЯ // Сибирский журнал науки и технологий – 2008. – С. 40-44.

[19] ГОСТ Р 55999-2014 Внутритрубное техническое диагностирование газопроводов. Общие требования. – 2015.02.01 – с.7.

[20] Методические указания по организации и исполнению программ диагностики промысловых трубопроводов. Компании ОАО «Газпром Нефть». – 2017. – с.43.

[21] Официальный сайт документации Python: сайт. – URL: https://www.python.org/ (дата обращения: 023.04.2020). – Текст: электронный.

[22] Официальный сайт документации C++: сайт. – URL: https://en.cppreference.com/w (дата обращения: 023.04.2020). – Текст: электронный.

[23] Официальный сайт документации React: сайт. – URL: https://ru.reactjs.org/ (дата обращения: 023.04.2020). – Текст: электронный.

[24] Официальный сайт документации Sqlite3: сайт. – URL: [https://www.sqlite.org/index.html (дата обращения: 023.04.2020). – Текст: электронный.

[25] Официальный сайт документации Scikit-Learn: сайт. – URL: https://scikit-learn.org/ (дата обращения: 023.04.2020). – Текст: электронный.

[26] Официальный сайт документации Pandas: сайт. – URL: https://pandas.pydata.org/ (дата обращения: 023.04.2020). – Текст: электронный.

[27] Официальный сайт документации Scipy: сайт. – URL: https://www.scipy.org/ (дата обращения: 023.04.2020). – Текст: электронный.

[28] Официальный сайт документации Flask: сайт. – URL: https://flask.palletsprojects.com/en/1.1.x// (дата обращения: 023.04.2020). – Текст: электронный.