Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико–Механический институт

Работа допущена к защите
Руководитель образовательной программы
_____ К.Н. Козлов
«_____» _____ 2023 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
РАБОТА БАКАЛАВРА
РАЗРАБОТКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ ФЕНОТИПА РАСТЕНИЙ НА ОСНОВЕ РАЗРЕЖЕННОГО РАЗЛОЖЕНИЯ ИСКУССТВЕННЫХ ИЗОБРАЖЕНИЙ, КОДИРУЮЩИХ ГЕНЕТИЧЕСКИЕ И ПОГОДНЫЕ ДАННЫЕ

по направлению подготовки 01.03.02 Прикладная математика и информатика

Направленность (профиль) 01.03.02_04 Биоинформатика

Выполнил
студент гр. 5030102/90401 Н.Н. Галлямова

Руководитель
доцент ВШПМиВФ, к.б.н. К.Н. Козлов

Консультант
по нормоконтролю Л.А. Арефьева

Санкт-Петербург
2023

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико–Механический институт

УТВЕРЖДАЮ
Руководитель образовательной программы
«Прикладная математика и информатика»
_____ К.Н. Козлов
«_____» _____ 2023г.

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

студенту Галлямовой Нэлли Наилевне гр. 5030102/90401

1. Тема работы: Разработка модели прогнозирования фенотипа растений на основе разреженного разложения искусственных изображений, кодирующих генетические и погодные данные.
2. Срок сдачи студентом законченной работы: июнь 2023 г.
3. Исходные данные по работе: Данные по времени цветения и однонуклеотидным полиморфизмам образцов дикого нута.
   Инструментальные средства:
   1. языки программирования Python
   2. среда разработки Jupyter Notebook, Google Colaboratory
   3. система контроля версий git

   Ключевые источники литературы:
   - Y. Bai, Z. Zhu, G. Jiang и H. Sun. «Blind Quality Assessment of Screen Content Images Via Macro-Micro Modeling of Tensor Domain Dictionary». IEEE Transactions on Multimedia, pp. 4259-4271, 2021 г. https://doi.org/10.1109/tmm.2020.3039382
   - M. Aharon, M. Elad and A. Bruckstein. «K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation». IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311-4322, 2006 г. https://doi.org/10.1109/tsp.2006.881199

4. Содержание работы:
    4.1. Введение. Обоснование актуальности
    4.2. Постановка задачи
    4.3. Разработка алгоритма
    4.4. Применение
    4.5. Результаты и их сравнительный анализ
    4.6. Выводы
    4.7. Заключение
5. Дата выдачи задания: 12.09.2022.

Руководитель ВКР _____К.Н. Козлов

Задание приняла к исполнению

Студент _____Н.Н. Галлямова

РЕФЕРАТ

На 47 с., 25 рисунков, 1 таблицу, 2 приложения

КЛЮЧЕВЫЕ СЛОВА: ПОБИТОВОЕ КОДИРОВАНИЕ, РАЗРЕЖЕННОЕ КОДИРОВАНИЕ, K-SVD АЛГОРИТМ, ОБУЧЕНИЕ СЛОВАРЯ, КАЧЕСТВО ИЗОБРАЖЕНИЯ, РЕГРЕССИОННАЯ МОДЕЛЬ.

Тема выпускной квалификационной работы: «Разработка модели прогнозирования фенотипа растений на основе разреженного разложения искусственных изображений, кодирующих генетические и погодные данные»

Данная работа посвящена разработке модели прогнозирования фенотипов растений с использованием генетических и климатических данных, представленных в виде искусственных изображений. Представлен новый оптимизированный алгоритм, который эффективно кодирует значения факторов в эти изображения. Кроме того, был разработан метод извлечения ключевых признаков и создания словарей признаков на основе алгоритма оценки качества изображения, используя K-SVD алгоритм обучения словаря и разреженного кодирования. Реализация этого подхода к представлению данных совмещена с машинным обучением, в частности, с машинами опорных векторов (SVM), для регрессионного анализа с целью прогнозирования фенотипов растений. Проведенные нами исследования позволили применить полученную систему на практике к прогнозированию времени цветения нута, что демонстрирует ее эффективность. В ходе работы был использован полный набор данных для построения и валидации модели, а также успешно проведен анализ факторов, наиболее существенно влияющих на точность модели. В результате, модель показала впечатляющие результаты в прогнозировании времени цветения, что может послужить основой для будущих исследований и разработок в области сельскохозяйственной геномики.

ABSTRACT

47 pages, 25 figures, 1 tables, 2 appendices

KEYWORDS: BITWISE CODING, SPARSE CODING, K-SVD ALGORITHM, DICTIONARY LEARNING, IMAGE QUALITY, PREDICTION MODEL.

The subject of the graduate qualification work is «Development of a plant phenotype forecasting model based on sparse decomposition of artificial images encoding genetic and weather data».

This work is dedicated to the development of a model for predicting plant phenotypes using genetic and climatic data, presented in the form of artificial images. A new optimized algorithm is introduced, which effectively encodes the values of factors into these images. Additionally, a method for extracting key features and creating feature dictionaries based on the image quality assessment algorithm, using the K-SVD dictionary learning and sparse coding algorithm, has been developed. The implementation of this data representation approach is combined with machine learning, specifically, with Support Vector Machines (SVM), for regression analysis with the aim of predicting plant phenotypes. Our research has allowed the application of the resulting system in practice for predicting the flowering time of chickpea, demonstrating its effectiveness. The study used a complete dataset for model construction and validation, and successfully conducted an analysis of the factors most significantly affecting the accuracy of the model. As a result, the model showed impressive results in predicting flowering time, which can serve as a basis for future research and development in the field of agricultural genomics.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние годы достижения в области технологий и сбора данных привели к большому количеству исследований в области изучения фенотипа растений. Данные признаки относятся к наблюдаемым физическим признакам, которые в свою очередь являются результатом взаимодействия генофонда и средой обитания. Одним из наиболее важных фенотипических признаков растений является время цветения.

Время цветения — это сложная характеристика, на которую влияют многие факторы, например, к ним относятся генетические факторы и факторы окружающей среды [1]. Понимание процесса их взаимодействия определяет сроки цветения, которые являются важной задачей в агрономии.

С генетической точки зрения несколько генов способствуют контролю времени цветения, либо стимулируя, либо подавляя данный процесс [2]. Каждый из этих генов кодирует белок, играющий определенную роль в сложном молекулярном механизме, регулирующем цветение. Генетический контроль времени цветения варьируется внутри вида среди разных сортов или штаммов.

Факторы окружающей среды, влияющие на время цветения, в основном включают температуру, продолжительность светового дня, количество выпавших осадков и количество солнечной радиации [3]. Например, некоторые растения зацветают только после определенного периода низких температур, процесс, известный как яровизация [4]. Другие цветут в ответ на изменение длины дня, что позволяет им синхронизировать свое цветение с определенным сезоном [5].

Однако взаимосвязь между генетическими факторами и факторами окружающей среды не является однозначной. Один и тот же генотип может демонстрировать разное время цветения в разных условиях — явление, известное как фенотипическая пластичность [6].

Сбор и анализ данных о генотипах растений и условиях окружающей среды может дать ценную информацию о механизмах, контролирующих время цветения. Это также может способствовать выведению сортов растений, которые лучше приспособлены к конкретному климату или условиям выращивания, что является важной задачей в контексте изменения климата [7] и растущего спроса на продукты питания [8].

Достижения в таких технологиях в сочетании со сложными статистическими и вычислительными методами, позволяют анализировать сложные генетические и

экологические факторы, определяющие время цветения. Изучение взаимосвязи данных факторов обеспечивает основу для разработки прогностических моделей. Эти модели можно использовать для прогнозирования времени цветения в различных условиях окружающей среды. Данное исследование имеет важное значение для селекции, сохранения растений и понимания адаптации и их эволюции.

Цель работы – разработка модели прогнозирования фенотипа растений по генетическим и погодным данным, которые представлены в виде искусственных изображений.

Задачи работы:

1. Разработка оптимизированного алгоритма кодирования значений факторов в искусственное изображение и методов извлечения характерных черт на основе создания словаря и разреженного кодирования.
2. Создание модели прогнозирования фенотипа растения на основе машины опорных векторов для регрессии.
3. Применение разработанных методов для разработки модели прогнозирования времени цветения нута по имеющемуся набору данных, выявление факторов, наиболее сильно влияющих на точность модели.

Таким образом, в данной работе будет предложена разработка алгоритма, направленного на создание модели прогнозирования, основанной на оценке качества искусственно сгенерированного изображения, полученного путем обработки начальных входных данных и впоследствии используемого для прогнозирования конечного результата.

ГЛАВА 1. ПОСТАНОВКА ЗАДАЧИ

Дана матрица входных данных $X \in \mathbb{R}^{k \times p}$.

Каждая строка $x_i \in \mathbb{R}^p$ представляет собой вектор факторов i-го растения. Здесь k - число растений, p - число погодных факторов и однонуклеотидных полиморфизмов, $i = 1, \ldots, k$.

Дан вектор $y \in \mathbb{R}^k$, где каждый элемент y_i представляет собой фенотип i го растения.

Необходимо найти функцию $F : X \to y$ такую, что минимизируется некоторая мера расхождения между прогнозированием модели и данными. Данную меру расхождения обозначим как $L(y, F(X))$, где L является некоторой функцией потерь.

Сформулируем задачу минимизации:

Найти функцию F^*, такую что:

$$F^* = \underset{F \in \mathcal{H}}{\operatorname{argmin}} L(y, F(X)),$$

где $\mathcal{H}$ - пространство гипотез

ГЛАВА 2. ОБЗОР СУЩЕСТВУЮЩИХ АЛГОРИТМОВ ОЦЕНКИ КАЧЕСТВА ИЗОБРАЖЕНИЙ

На данный момент существует два типа алгоритмов, созданных для оценки качества изображения, а именно алгоритмы Image Quality Assessment (IQA) и алгоритмы слепой оценки качества изображения – Blind Image Quality Assessment (BIQA) [9].

IQA включает в себя алгоритмы, которые оценивают качество на основе сравнения исходного изображения или еще известного, как эталонного изображения и обработанного изображения. Основная цель состоит в том, чтобы измерить сходство или различие между двумя изображениями.

Есть два вида IQA: Full Reference (FR) и Reduced Reference (RR) [10]. С первым типом FR-IQA доступна вся информация об эталонном изображении, тогда как RR-IQA позволяет использовать только часть этих данных.

Также существует алгоритмы слепой оценки качества изображения (BIQA), они работают без эталонного изображения. Задача оценки качества изображения основывается исключительно на обработанном изображении, что повышает сложность ее решения. Они также известны как методы IQA без эталона (NR). Алгоритмы BIQA стремятся имитировать способность зрительной системы человека определять качество без необходимости в эталоне.

Каждый из этих типов алгоритмов имеет свои достоинства и ограничения, и выбор между ними обычно основывается на конкретных потребностях рассматриваемой задачи.

Изначально, в целях поиска необходимого нам алгоритма оценки были проанализированы множество существующих, информация о которых предоставлена в Таблице 2.1.

Таблица 2.1

Характеристики алгоритмов оценки качества изображений

Порядковый номер работы	Искусственное изображение	Оценка качества без эталона	Регрессионная модель
1 [9]	+	+	+
2 [10]	+	+	+
3 [11]	-	-	-
4 [12]	+	-	-
5 [13]	+	-	-
6 [14]	+	-	-
7 [15]	+	-	-
8 [16]	+	-	-
9 [17]	-	+	-
10 [18]	-	+	+
11 [19]	-	+	+
12 [20]	+	+	+
13 [21]	+	-	-
14 [22]	+	-	-
15 [23]	-	-	-
16 [24]	+	-	-
17 [25]	-	+	-
18 [26]	+	+	+

Таблица 2.1

Характеристики алгоритмов оценки качества изображений (продолжение)

Порядковый номер работы	Искусственное изображение	Оценка качества без эталона	Регрессионная модель
19 [27]	+	+	+
20 [28]	-	-	-

Однако среди данного множества алгоритмов выбор был сделан в пользу алгоритма, предложенного в статье [10], он основан на разряженном разложении искусственных изображений. Это решение было принято на основе сравнительного анализа характеристик данного алгоритма, демонстрирующего наилучшее сочетание необходимых признаков, что делает его подходящим для решения поставленной конкретной проблемы – прогнозирования фенотипа растений.

Данный алгоритм использует метод слепой оценки качества изображения, что является его несомненным достоинством в нашем случае, поскольку у нас нет доступа к эталонным изображениям и мы не знаем, что такое эталон в контексте нашей задачи. Также преимуществами данного алгоритма является его совместимая работа с искусственными изображениями и использование регрессионной модели для прогнозирования качества изображения.

ГЛАВА 3. РАЗРАБОТКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ

3.1. Алгоритм генерации искусственных изображений

Перейдем к рассмотрению алгоритма, генерирующего искусственные изображения. Данный этап работы преобразует исходные входные численные данные в пиксельные значения, с помощью которых далее компонуются изображения, которые впоследствии применяются в алгоритме оценки качества получившихся новых наборов данных.

Приведем более подробно описание последовательных шагов каждого этапа.

1. Преобразование численных значений в цветовые значения в промежутке от 0 до 255.

   1. Кодирование генетической информации.

   В основе данной части работы алгоритма лежит – битовое кодирование информации. Для простоты работы алгоритма сделаем следующую замену:

   1) Доминантная гомозигота «АА» заменяется численным значением – «2».
   2) Рецессивная гомозигота «аа» заменяется численным значением – «0».
   3) Гетерозигота «Аа» заменяется численным значением – «1».

   Основная структура кодирования информации представлена на рис.3.1.
   Резюмируя, получаем, что последовательность ОНП кодируется с помощью использования битового сдвига для кодирования информации. Каждый бит сдвигается влево на указанное количество позиций, а именно на номер индекса ОНП в исходной последовательности, а затем применяется операция «ИЛИ» (‘|’) или же, как она еще известна, операция «установки» бита с накапливающей значение переменной.
   Результатом являются два числа, в которых биты представляют информацию о кодируемых значениях, далее, два этих значения заполняют двумерную матрицу соответствующей размерности и канала соответствующей аллели.

   2. Кодирование климатических данных.

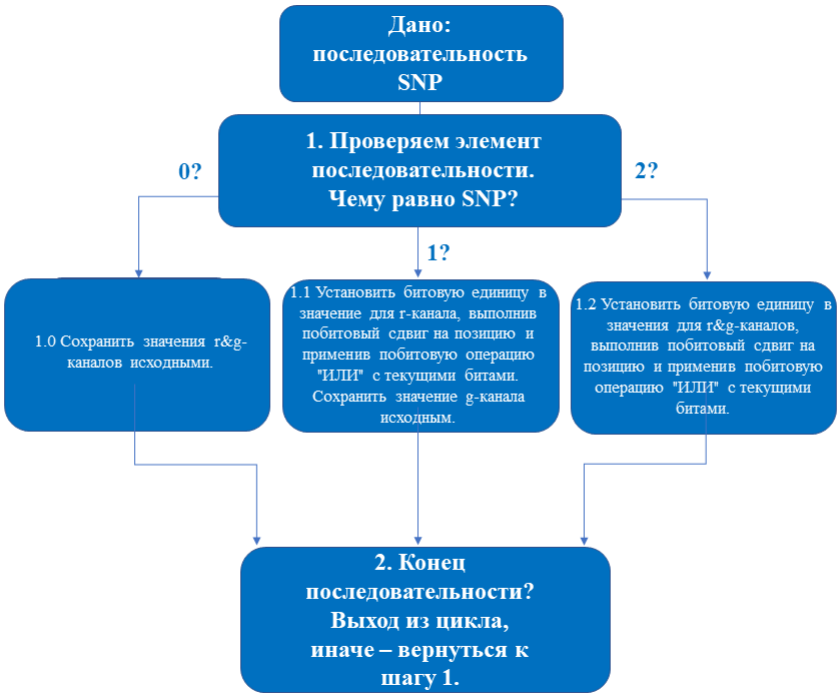

Рис.3.1. Блок-схема работы алгоритма генерации искусственных изображений.

Данная информация кодируется следующим образом:

Input data - int number

return [number // 256, number % 256]

Таким образом, в данном случае, вся численная информация о погоде, имеющаяся за конкретный наблюдаемый день, преобразовывается в соответствии с алгоритмом, представленным выше, и записывается в один столбец матрицы последовательно. Первый элемент относится к одному цветовому каналу, второй элемент к другому цветовому каналу, для каждого канала своя матрица данных.

2. Комбинирование полученных матриц данных в итоговое изображение. Заключительным этапом формирования изображения является их общая компоновка. Данный шаг подробно представлен на рис.3.2.

   Дадим последовательное описание каждого шага, представленного на ней.

   1. Каждая матрица данных, кодирующая определенную информацию, отвечает за соответствующий ей только один из RGBA–каналов.
   2. Далее все 4 канала образуют итоговое изображение.
   3. На следующем этапе происходит дублирование изображения.

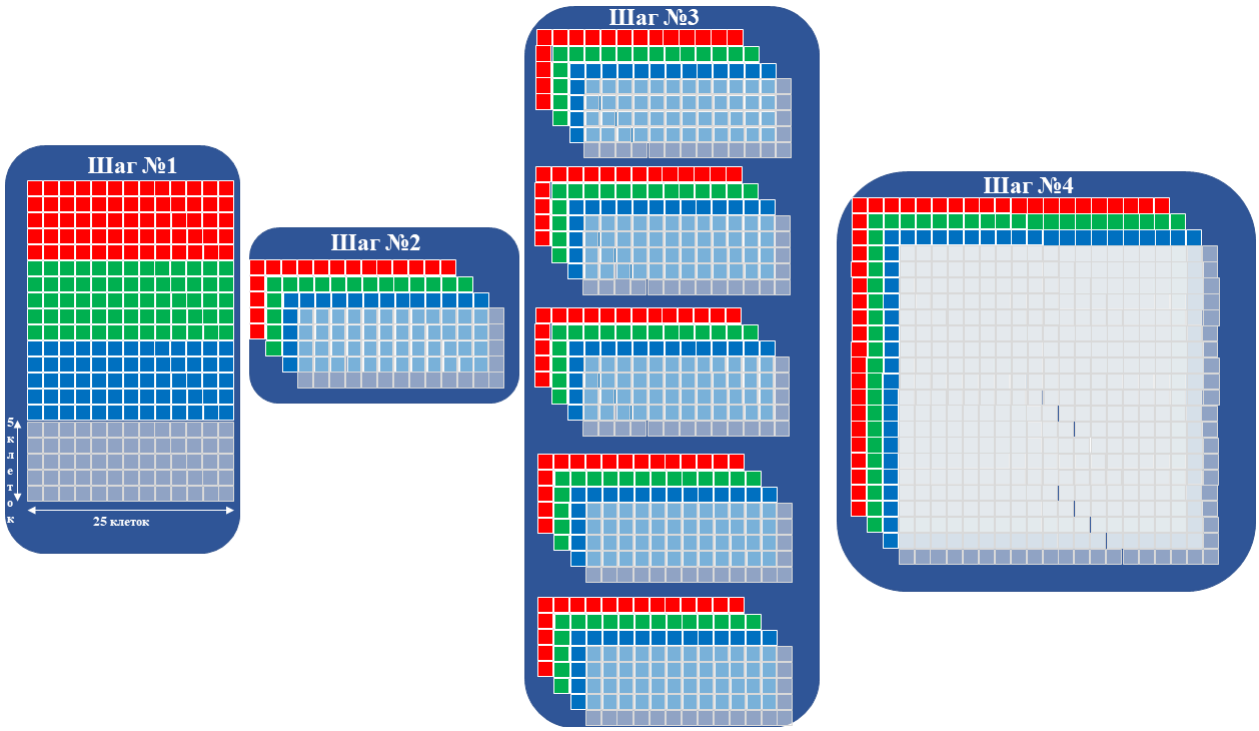

Рис.3.2. Блок-схема комбинации цветовых каналов изображения.

4. Дубликаты сливаются в одну колонку и образуют итоговое изображение. Цель данного этапа - повысить размерность исходных данных, тем самым увеличив последующую эффективность разработанной модели.

Итогом работы алгоритма является изображение в режиме «RGBA» со своими вычисленными значениями в каждой из ячеек данных матриц цветового канала.

3.2. Алгоритм оценки качества изображений

Данный алгоритм основывается на разреженном представлении, в котором общие векторы признаки – это комбинация микро- и макропризнаков. Структура предлагаемого метода продемонстрирована на блок-схеме в Приложении 1. Предлагаемый метод включает два этапа:

1. Обучение словаря.
2. Контроль качества.

Сам словарь строится на основе главных компонент после применения тензорного разложения Такера. Для прогнозирования качества находятся микро- и макропризнаки тестового искусственного изображения с помощью обученно-

го словаря и модели выделения признаков, далее реализуется SVR-модель для прогнозирования оценки качества.

Выделим четыре основных шага разработки данного алгоритма, предложенных на блок-схеме.

Шаг №1. Подготовка данных.

1. Подготовка данных – загрузка изображений.

2-3. Предварительная обработка – тензорное разложение Такера.

Тензорное разложение Такера [10] – форма анализа главных компонент высшего порядка. Происходит декомпозиция тензора $\chi \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ в тензор-ядро $\varsigma \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_N}$, умноженный на группу фактор-матриц $\mathbf{Y}^{(n)} \in \mathbb{R}^{I_N \times J_N} (1 \leqslant n \leqslant N)$

В RGBA данные представлены тензором четвертого порядка

$$\chi \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4},$$

где I_1, I_2, I_3 и I_4 указывают размеры красного, зеленого, синего и альфа каналов изображения RGBA, соответственно.

Тогда разложение выглядит следующим образом:

$$\chi \approx \varsigma \times \mathbf{Y}^{(1)} \times \mathbf{Y}^{(2)} \times \mathbf{Y}^{(3)} \times \mathbf{Y}^{(4)},$$

где $\varsigma \in \mathbb{R}^{I_1 \times J_2 \times J_3 \times J_4}$ - ядро, $\mathbf{Y}^{(1)} \in \mathbb{R}^{I_1 \times I_1}$, $\mathbf{Y}^{(2)} \in \mathbb{R}^{I_2 \times I_2}$, $\mathbf{Y}^{(3)} \in \mathbb{R}^{I_3 \times I_3}$ и $\mathbf{Y}^{(4)} \in \mathbb{R}^{I_4 \times I_4}$ – фактор-матрицы, чаще всего они ортогональны, каждая может быть использована в качестве основной компоненты.

В алгоритме будет использоваться следующая компонента – $\mathbf{Y}^{(2)}$, так как она среди всех имеющихся представляет больше информации, как по яркости, так и по цветовой насыщенности.

4. Слияние изображений воедино.

Для того, чтобы использовать весь набор изображений, необходимо объединить все изображения в один большой набор данных. Это позволяет алгоритму проанализировать все доступные признаки из всех изображений и использовать эту информацию эффективно. Объединение данных также способствует более тесной интеграции признаков из различных изображений, что в свою очередь может улучшить обучение и прогнозирование.

Шаг №2. Обучение словаря.

5. Обучение словаря.

Для обучения словаря используется алгоритм K-SVD [29].

Основная идея алгоритма заключается в следующем:

Каждый блок изображения может быть представлен в виде взвешенной суммы изображений-шаблонов (так называемых атомов), хранящихся в заранее подготовленном словаре.

На этапе обучения изображения-шаблоны получаются на основе набора блоков реальных изображений, причем метод k-ближайших соседей используется для поиска похожих шаблонов, а метод сингулярных разложений (SVD) используется для ускорения сходимости итерационного процесса составления словаря.

K-SVD алгоритм.

Задача: найти наилучший словарь, который представляет входной сигнал как композицию разреженных представлений, решая задачу минимизации

$$\min_{\mathbf{D},\mathbf{X}} \left\{ \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_{\mathrm{F}}^2 \right\} \forall i, \|x_i\| \leqslant T_0,$$

то есть количество ненулевых значений может быть больше 1, но меньше какого-то фиксированного числа T_0, которое задается.

Псевдокод K-SVD алгоритма представлен на рис.3.3.

Рассмотрим подробно процесс разреженного кодирования.

Для того, чтобы вычислить разреженное представление сигнала воспользуемся алгоритмом ортогонального согласованного преследования (Orthogonal Matching Pursuit - OMP), представленным на рис.3.4.

Шаг №3. Обучение модели оценки качества изображения.

5-6. Разреженное представление.

Данные для обучения модели представляются в виде разреженного представления с использованием обученного словаря, то есть данные из исходного пространства признаков преобразуются в пространство целевого словаря. Разреженное кодирование используется для получения желаемого представления признаков, так как оно представляет входной сигнал как линейную комбинацию атомов в словаре.

Здесь мы используем $\hat{\mathbf{Y}} = \{y_i\}_{i=1}^N$ - основную компоненту после использования тензорного разложения, затем разреженное представление x_l может быть вычислено для произвольного патча с помощью обученного словаря $\mathbf{D} \in \mathbb{R}^{p \times K}$ с помощью использования алгоритма OMP следующим образом:

$$\hat{\mathbf{X}} = \operatorname{argmin} \|\hat{x}_i\|_0, \text{ так что } \|\hat{\mathbf{Y}} - \mathbf{D}\hat{\mathbf{X}}\|_2^2 \leqslant T,$$

где T - предопределенная пороговая ошибка и $\hat{\mathbf{X}} = [x_1, \ldots, x_N] \in \mathbb{R}^{p \times N}$.

7-9. Выделение признаков.

1.**Procedure** `Разреженное кодирование`(Y, $D^{(J-1)}$)
    Input: Входные данные Y, словарь $D^{(J-1)}$
    Output: Разреженное представление X
2. Инициализация: X
3. Вычисление разреженного представления X с использованием $D^{(J-1)}$
4. **return** X
5.**Procedure** `Обновление словаря`($D^{(J-1)}$, X)
    Input: Словарь $D^{(J-1)}$, разреженное представление X
    Output: Обновленный словарь $D^{(J)}$
6. Инициализация: $D^{(J)} \leftarrow D^{(J-1)}$
7. **for** $k = 1$ **to** K **do**
8. Определить индексы $\omega_k = \{i | 1 \leqslant i \leqslant N, x_T^k(i) \neq 0\}$
9. Вычислить матрицу ошибок $E_k = Y - \sum_{j \neq k} d_j x_T^j$
10. Ограничить E_k, выбирая только те номера столбцов, которые соответствуют ω_k, получая E_k^R
11. Применить сингулярное разложение $E_k^R = U \Delta V^T$
12. Обновить d_k как первый столбец U
13. Обновить x_T^k как произведение первого столбца V на $\Delta(1, 1)$
14. **end**
15. **return** $D^{(J)}$
16.**Procedure** `Разреженное кодирование и обновление словаря`(Y, $D^{(0)}$)
    Input: Входные данные Y, начальный словарь $D^{(0)}$
    Output: Обновленный словарь D
17. Инициализация: $D \leftarrow D^{(0)}$, $J \leftarrow 1$, Условие сходимости $\leftarrow$ False
18. **while** *не выполнено условие сходимости* **do**
19. $X \leftarrow$ Разреженное кодирование(Y, $D^{(J-1)}$)
20. $D^{(J)} \leftarrow$ Обновление словаря($D^{(J-1)}$, X)
21. $J \leftarrow J + 1$
22. **if** *условие сходимости выполнено* **then**
23. Условие сходимости $\leftarrow$ True
24. **end**
25. **end**
26. **return** D
    Input: Входные данные Y, начальный словарь $D^{(0)}$
    Output: Обновленный словарь D
27.$D \leftarrow$ `Разреженное кодирование и обновление словаря`(Y, $D^{(0)}$)

Рис.3.3. Алгоритм разреженного кодирования и обновления словаря.

1.**Procedure** Разреженное кодирование(D, x, e, max_iter)
Input: Словарь D, сигнал x, условие остановки (точность e или максимальное количество итераций max_iter)
Output: Разреженное представление $\gamma : x \approx D\gamma$
2. Инициализация: $I \leftarrow \varnothing$, $r \leftarrow x$, $\gamma \leftarrow 0$
3. **while** *не выполнено условие остановки* **do**
4. $k^* \leftarrow \arg\max_k |\langle d_k, r\rangle|$ $I \leftarrow I \cup \{k^*\}$ Находим решение для $y = Dx$: $\gamma_I \leftarrow (D_I)^+ x$ Вычисляем ошибку: $r \leftarrow x - D_I\gamma_I$
5. **end**
6. **return** γ

Рис.3.4. Orthogonal Matching Pursuit.

После применения разреженного представления векторы признаков для всех патчей рассчитываются из тестового изображения, данные векторы отражают ключевые характеристики входного изображения в разреженном представлении. В пространстве словаря - векторы признаков, извлеченные из отдельных фрагментов (патчей) изображения, должны быть объединены в конечный вектор признаков всего изображения, удобный для регрессии качества, так как целью оценки качества изображения (IQA) является получение единой оценки изображения. Далее, полученные вектора признаков были проанализированы на уровне каждого патча, учитывая микро- и макро детали изображения. Таким образом, была создана модель, которая объясняет, как данные признаки формируются, используя принципы статистики (закон больших чисел Бернулли) и разреженного кодирования. Далее подробно опишем процесс создания модели.

1) Получение микро признаков.

Все атомы обученного словаря используются как основные элементы для описания изображения. Для словаря $\mathbf{D} \in \mathbb{R}^{p\times k}$, тестовое изображение может быть представлено с помощью разреженного кодирования следующим образом:

$$\hat{\mathbf{Y}} = \mathbf{D}\hat{\mathbf{X}} = \sum_{i=1}^{N} \mathbf{D}\hat{x}_i \Rightarrow \sum_{i=1}^{N}\sum_{j=1}^{k} \alpha_{i,j} d_j,$$

где $d_j \in \mathbb{R}^p$ обозначает j-ый атом обученного словаря, и каждый атом это p-мерный вектор, $\alpha_{i,j}$ - обозначает соответствующий разреженный коэффициент j-го атома для i-го патча.

В статье [10] показано, что оптимальное использование всех векторов признаков основывается на новом подходе, использующем логнормальное распределение. Основная его суть - сосредоточиться на ненулевых коэффициентах, так как они являются наиболее значимыми для восстановления изображения. Статистический анализ показал, что распределение этих коэффициентов для каждого атома (базового элемента словаря) обычно следует логнормальному распределению со следующей функцией плотности вероятности:

$$f(x;\mu;\sigma) = \frac{1}{\sqrt{2\pi}x\sigma} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}$$

$$M[x] = e^{\mu+\frac{\sigma^2}{2}}$$

Предложенная схема объединения вектор признаков работает следующим образом: на вход подается матрица $\hat{\mathbf{X}} = [x_1, \ldots, x_N] \in \mathbb{R}^{p\times N}$, затем из нее извлекаются только те коэффициенты, которые не равны нулю, данный процесс происходит для каждого атома отдельно. То есть, иными словами, для каждого атома (базового элемента) рассматриваются только те части изображения, где этот атом имеет ненулевой коэффициент. Далее, после этого берутся абсолютные значения этих ненулевых коэффициентов и используются для вычисления среднего значения и стандартного отклонения. Математически это выражается следующим образом

$$f^{\text{mic}} = e^{\mu+\frac{\sigma^2}{2}},$$

где $f^{\text{mic}} = \left[f_1^{mic}, \ldots, f_k^{\text{mic}}\right]$ представляют векторы микропризнаков для всех атомов.

2) Получение макропризнаков.

Микропризнаки характеризуются с помощью объединения ненулевых значений коэффициентов атомов, то есть интенсивности для каждого атома, тогда как состав его количества может отражать макропризнаки. Согласно закону Бернулли больших чисел, для случайного события A в случайной среде, частота случайного события A в рамках большого набора независимых испытаний приближается к вероятности этого события.

Теорема Бернулли:

μ_n - число наступлений события A в n независимых испытаниях, p - вероятность наступления события A в одном испытании $\Rightarrow \forall\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

В нашем случае событие A - появление конкретного «атома». При анализе большого количества искаженных изображений для изначального одного изображения появление атома будет зависеть от типа и интенсивности искажения. То есть, если сосредоточиться на определенном «атоме», относительная частота его появления в искаженных изображениях будет стремиться к его истинной вероятности появления. Данное знание позволяет получить более точное представление о влиянии искажений на качество изображений.

Также можно оценить важность конкретного атома, вычисляя количество его появлений среди всех атомов, данное количество преобразуется в вероятность появления этого атома с помощью процесса нормализации.

$$f_i^{mac} = \frac{n_i}{\sum_{i=1}^{k} n_i}$$

где $f^{mac} = \left[f_1^{mac}, \ldots, f_k^{mac}\right]$ представляют векторы макропризнаков для всех атомов, n_i- количество появлений атома i для тестового изображения.

Таким образом, конечный вектор признаков изображения представляет собой объединение двух векторов признаков и обозначается как

$$f = \left[f^{mic}, f^{mac}\right] \in \mathbb{R}^{2k\times 1}$$

10-11. Обучение SVR-модели.

С помощью метода SVR, будем реализовывать прогнозирование оценки качества изображения, обозначаемой как Q, используя найденные вектора макро- и микропризнаков $\left[f^{mic}, f^{mac}\right]$, а также нелинейную функцию f. Таким образом, оценка качества определяется как $Q = f\left(f^{\text{mic}}, f^{\text{mac}}\right)$, где f находится с помощью SVR. Здесь $f : \mathbb{R}^{2k\times 1} \rightarrow \mathbb{R}$ обучается следующим образом:

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) K\left(x_i, x\right) + b$$

где α и α^*- множители Лагранжа, $b-$ параметр смещения, $K\left(x_i, x\right)$ − ядро радиальной базисной функции (rbf-ядро), l - количество обучающих примеров.

Формула для вычисления rbf-ядра:

$$\exp\left(-\gamma \left\|x - x'\right\|^2\right), \gamma > 0$$

Обучение происходит путем минимизации функции потерь, которая измеряет разницу между спрогнозированным значением функции f и реальными значениями.

То есть необходимо решить следующую задачу минимизации:

$$\min_{w,b,\xi,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right)$$

при условиях:

1. $y_i - w \cdot \varphi\left(x_i\right) - b \leqslant \varepsilon + \xi_i$

2. $w \cdot \varphi\left(x_i\right) + b - y_i \leqslant \varepsilon + \xi_i^*$

3. $\xi_i \geqslant 0, \xi_i^* \geqslant 0 \forall i$

Переменные ξ, ξ^* показывают, насколько можно допустить отклонение от параметра ε, параметр C - параметр регуляризации, контролирует, насколько сильно можно избегать больших отклонений, то есть контролирует баланс между сложностью модели и степенью допустимой ошибки, w - вектор весов, $\varphi(x)$ - преобразование, применяемое к входным данным, b - смещение.

Рассмотрим также применение различных методов для оптимизации SVR-модели.

1. Настройка гипер-параметров модели оценки.

Метод GridSearchCV из библиотеки Scikit-learn [29], как уже ясно из названия, производит поиск по сетке для заданных параметров поиска со всеми возможными комбинациями и выводит наилучшую модель по заданной метрике.

2. Кросс-валидация.

Для кросс-валидации будет использоваться подход - K-Fold CV [30]. Обучающий набор данных разделяется на k меньших подмножеств и далее выполняется следующая процедура для каждого фолда:

1) Модель обучается, используя $k - 1$ фолдов как обучающие данные;

2) Полученная модель проверяется на оставшейся части данных, то есть та часть, которая осталась, используется в качестве тестового набора данных для расчета точности модели.

Результирующий итог - показатели метрики модели, отраженные в каждом из фолдов, суммируются и находится среднее значение.

3. Метрики качества.

Были выбраны следующие метрики качества:

1) Maximum Absolute Error

Метрика, позволяющая измерить максимальное отклонение спрогнозированного значения от истинного значения среди всех данных. Вычисляется по следующей формуле:

$$MaxError = \max \left| y_{\text{true}} - y_{\text{pred}} \right|$$

2) Mean Absolute Error

Метрика, позволяющая вычислить среднее значение разностей между спрогнозированными значениями и истинными. Данная метрика проста в использовании и в целом полезна, поскольку дает представление о том, насколько наша модель в среднем отклоняется от истинных значений.

Вычисляется по следующей формуле:

$$MAE = \frac{1}{N} \Sigma \left| y_{\text{true}} - y_{\text{pred}} \right|,$$

где N - общее количество данных.

3) Root Mean Squared Error

Метрика, позволяющая вычислить квадратный корень из среднего квадратов разниц между спрогнозированными и истинными значениями. MSE штрафует наибольшие ошибки сильнее, чем наименьшие, то есть она полезна, если в приоритете хочется учитывать большие ошибки.

Вычисляется по следующей формуле:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(y_{\text{true}}^{(i)} - y_{\text{pred}}^{(i)} \right)^2,$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_{\text{true}}^{(i)} - y_{\text{pred}}^{(i)} \right)^2},$$

где N - общее количество данных.

11. Итоговый результат - обученная SVR-модель используется для оценки качества изображения.

Для разработки алгоритма были использованы следующие программные и аппаратные средства:

1. Программное обеспечение.

1.1) Язык программирования - Python (v3.9);

1.2) Библиотеки - Pandas (для обработки и анализа входных данных), Numpy (для численных вычислений), Scikit-learn (основная библиотека, используется для

обучения SVR модели и для алгоритма обучения словаря), Matplotlib и Seaborn (используются для визуализации);

1.3) Среды разработок - Google Colaboratory, Jupyter Notebook;

1.4) Система контроля версий - git.

2. Аппаратное обеспечение.

2.1) Процессор - Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz1.99 GHz;

2.2) Память - 16,0 ГБ.

ГЛАВА 4. РЕЗУЛЬТАТЫ

В данном разделе представлены ряд результатов и анализов, полученные с помощью разработанной модели прогнозирования фенотипа растений, а также описание набора исходных данных.

Будет продемонстрирована последовательная работа алгоритма, а затем проанализированы его результаты с целью изучить качество модели и зависимость ее от исходных параметров и полученных признаков.

4.1. Описание набора данных

Данные представляют собой генетические и климатические данные для каждого растения и наблюдения за погодными данными, где данное растение было выращено, соответственно. В файле с генетической информацией содержится необходимая нам следующая информация:

1. Растение и информация о каждом его ОНП.

В файле с погодными данными содержится необходимая нам следующая ин-формация:

1. Значения температур - максимальная и минимальная температуры в течение дня.
2. Значение солнечной радиации в течение дня.
3. Длительность светового дня.
4. Количество выпавших осадков в течение дня.

Данные о погоде брались следующим образом: за 5 дней до дня цветения и 20 дней после.

Также продемонстрируем для наглядности гистограмму распределений значений фенотипа в наборе данных (рис.4.1).

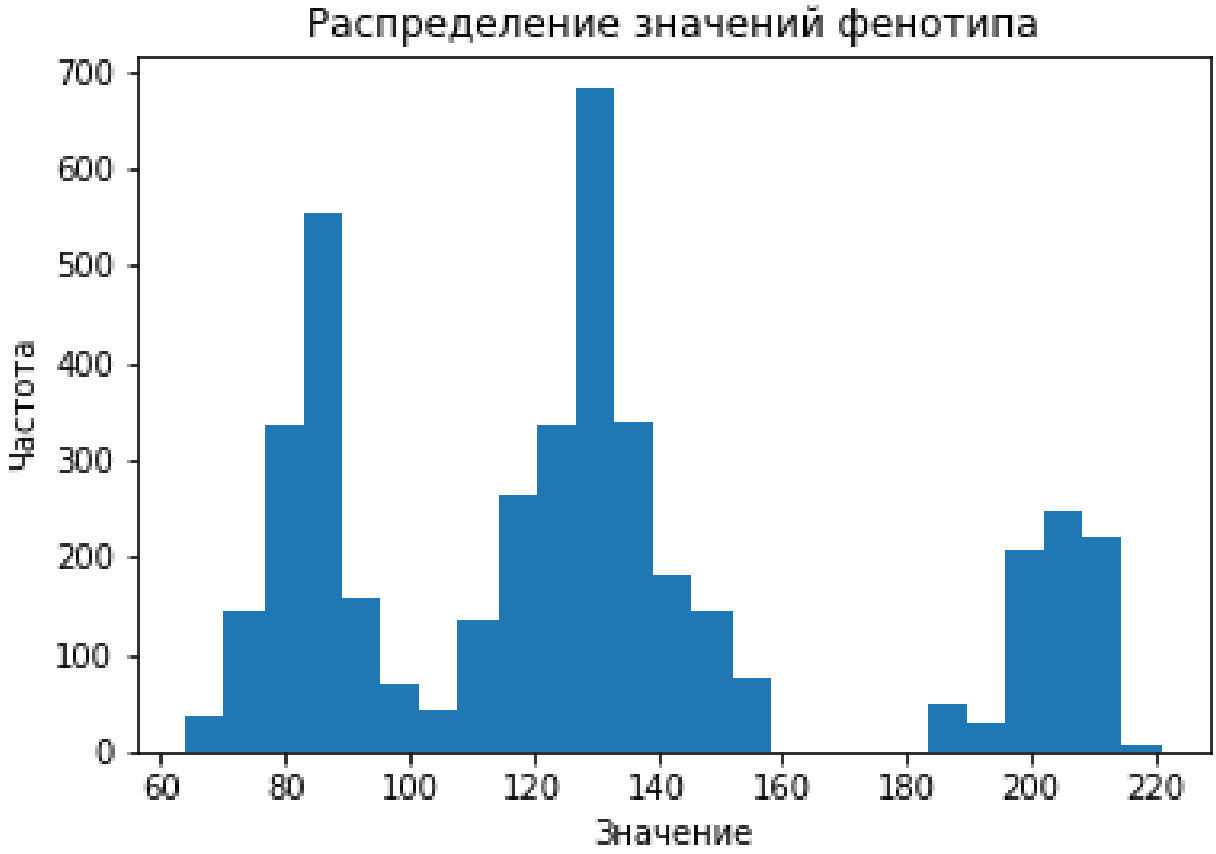

Рис.4.1. Распределение значений фенотипов в наборе данных.

По ней заметно, что значения в целом распределены примерно по трем промежуткам, если первые два относительно сбалансированы между собой, то третий не балансируют со всеми остальными. В дальнейшем это может сказаться на последующих результатах обучения модели, так, например, модель может оказаться смещенной в сторону данных с большим количеством экземпляров, то есть это может значить, что модель будет хуже прогнозировать результаты с данными, которых меньше.

4.2. Алгоритм генерации искусственных изображений

На начальном этапе разработки модели прогнозирования используется алгоритм генерации искусственных изображений. Рассмотрим один из примеров изображения (рис.4.2), сгенерированного с помощью алгоритма битового кодирова-

ния информации. Его представление по каналам в градациях серого представлено на рис.4.3.

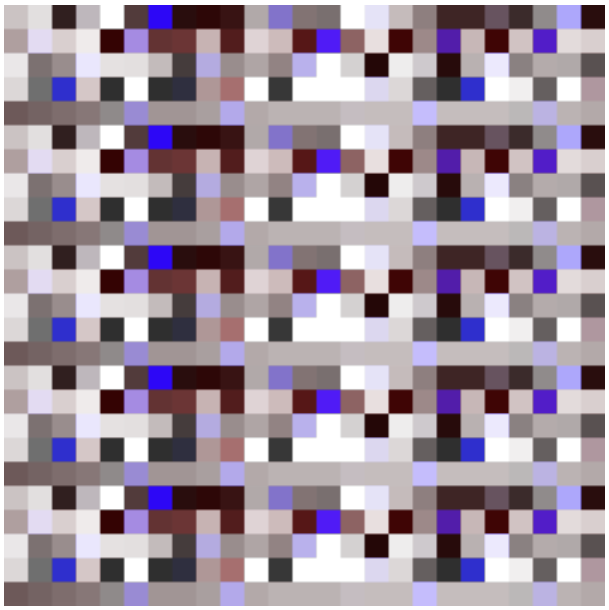

Рис.4.2. Пример изображения №1.

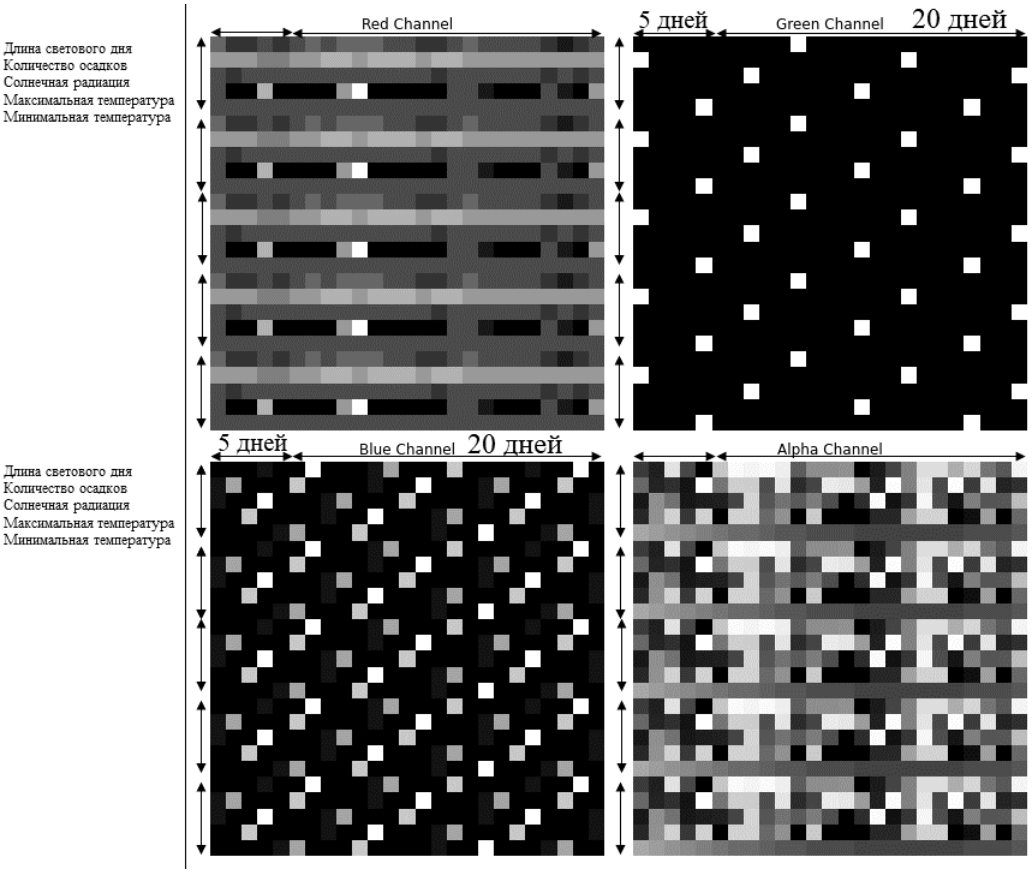

Рис.4.3. Представление искусственного изображения по каналам в градациях серого.

Ниже также продемонстрированы (рис.4.4) другие примеры сгенерированных изображений, демонстрирующие работу алгоритма.

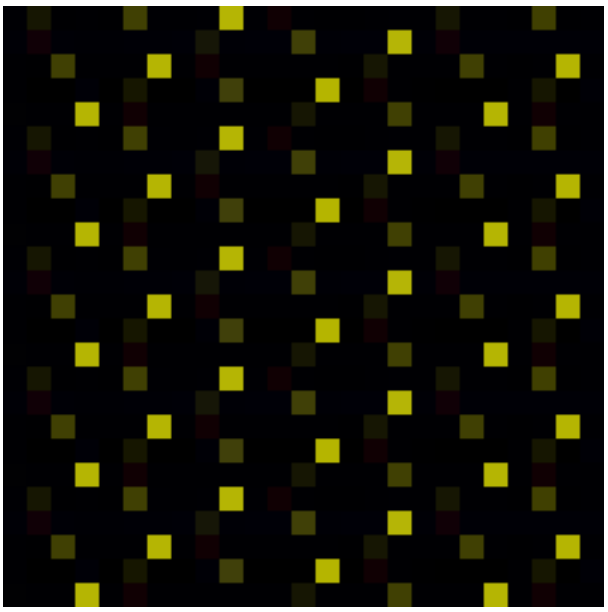

Рис.4.4. Пример изображения №2.

Таким образом, создание искусственно сгенерированных изображений на основе исходного набора данных предоставило нам разнообразный набор новых изображений, что позволяет в итоге в дальнейшем работать с ними, определяя в них существенно важные признаки для работы с итоговой моделью.

4.3. Алгоритм оценки качества изображений

1) Слияние изображений воедино.
   Для обучения словаря были случайным образом выбраны 75 искусственно сгенерированных изображений, которые последовательно сливались в одну строку.
   Таким образом, учитывая размерность изображений и количество выбранных изображений для обучения словаря получается матрица размером 625×75. В каждой строке этой матрицы представлено одно из 75 изображений, каждое изначально имеющее размерность 25×25. Изображение вытягивается в одну строку, то есть имеет размерность 625×1. При этом значения пикселей остаются теми же, которые содержались изначально в исходных искусственно сгенерированных изображениях.
   Итоговый результат "слитого"изображения (рис.4.5).
2) Предварительная обработка – тензорное разложение Такера.

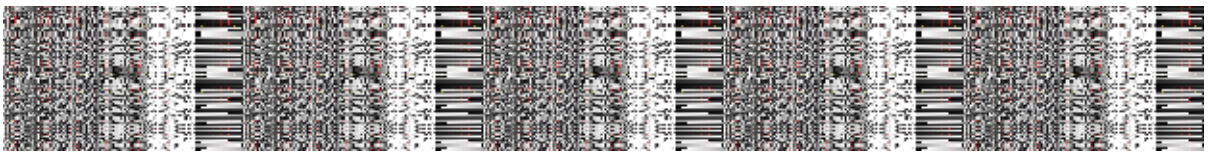

Рис.4.5. Пример "слитого"изображения.

Далее для всего набора сгенерированных изображений и для «слитого» изображения применялось тензорное разложение Такера.
Для одного изображения поэтапное применение разложения представлено ниже (рис.4.6, 4.7, 4.8).

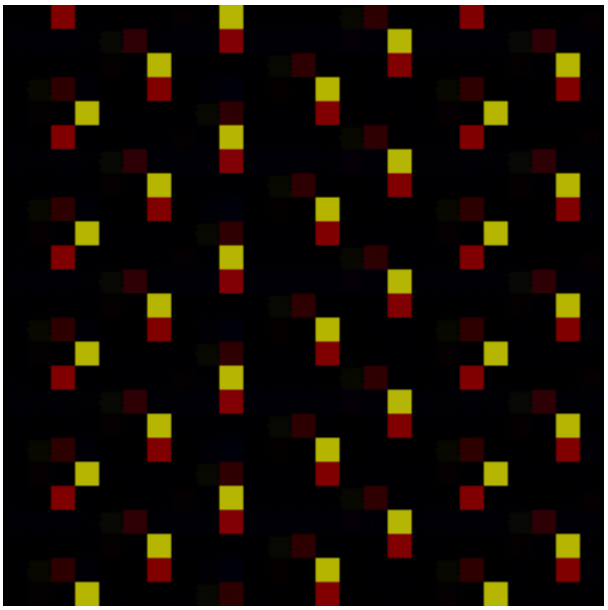

Рис.4.6. Исходное изображение.

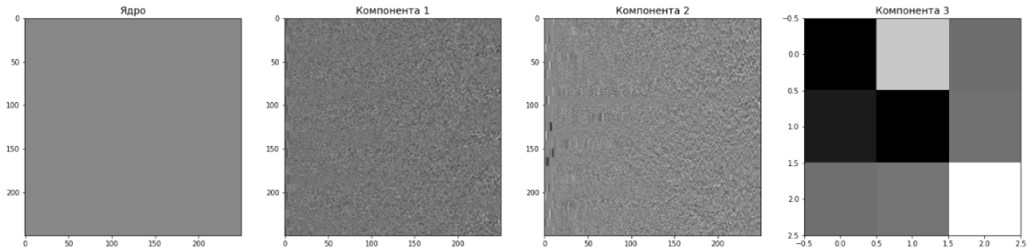

Рис.4.7. Тензорное разложение Такера.

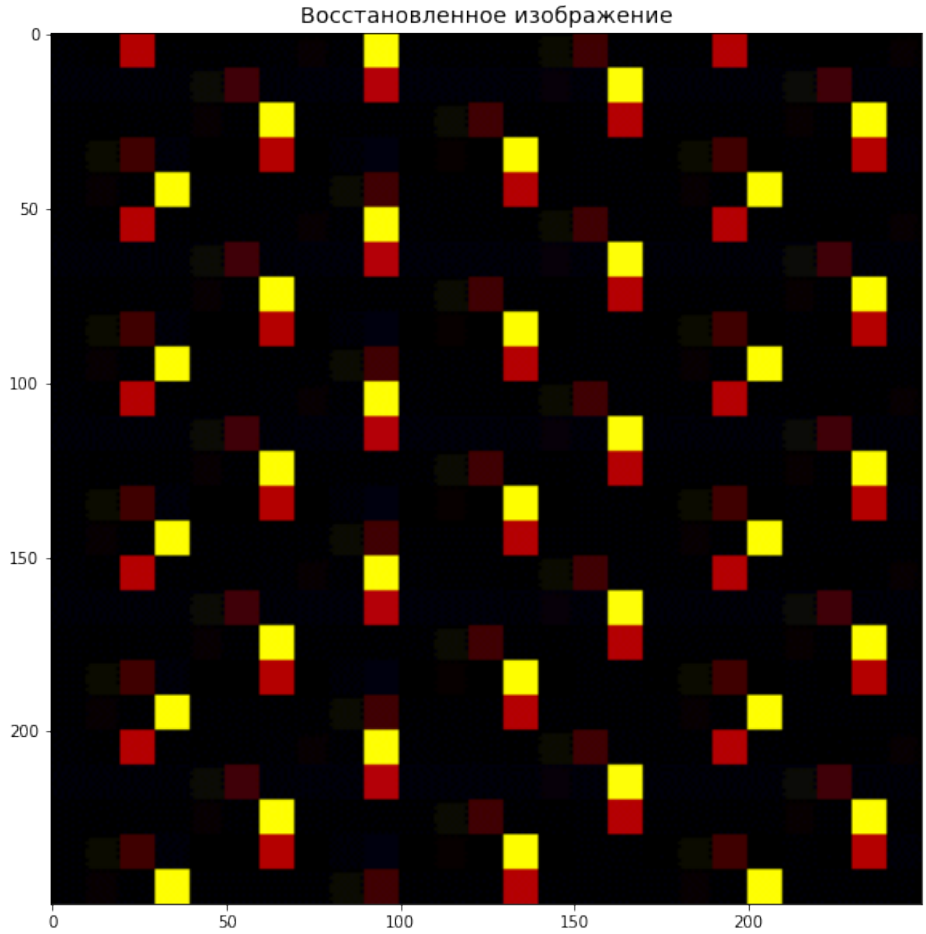

Рис.4.8. Восстановленное изображение.

Для «слитого» изображения получим следующее разложение (рис.4.9, 4.10, 4.11).

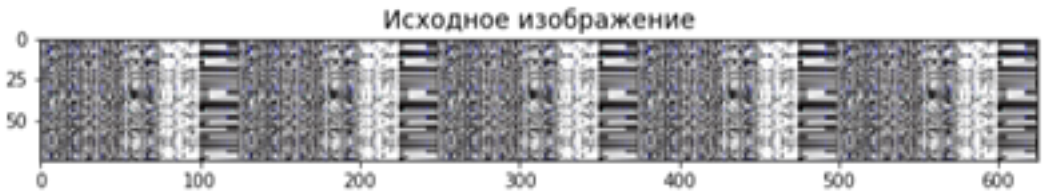

Рис.4.9. Исходное изображение.

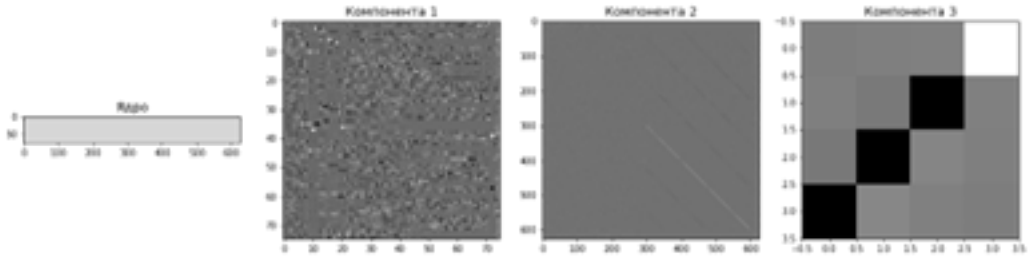

Рис.4.10. Тензорное разложение Такера.

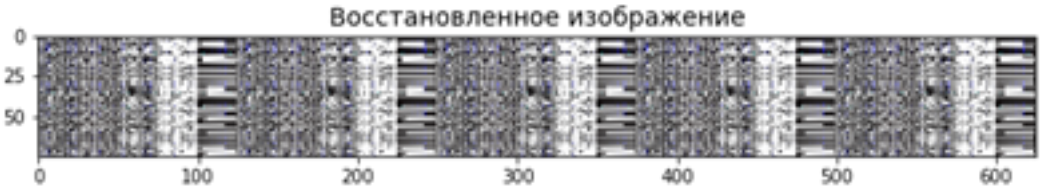

Рис.4.11. Восстановление изображение.

Стоит заметить, что в обоих случаях восстановленные изображения близки к исходным.

3) Извлечение патчей.

   Для обучения словаря «слитое» изображение разбивалось на патчи размером 4×4. Всего из изображения 625×75 было получено 44784 патча. Патчи извлекались с перекрытием, которое составляло 1 пиксель.

   1. В каждом ряду изображения можно взять $(625 - 4 + 1) = 622$ патча.
   2. В каждом столбце изображения можно взять $(75 - 4 + 1) = 72$ патча.

   Таким образом, общее количество патчей будет $622 \times 72 = 44784$ патча.

   Продемонстрируем первые 36 патчей из данного набора патчей (рис.4.12).

4) Проверка гипотезы о логнормальном распределении числа ненулевых коэффициентов.

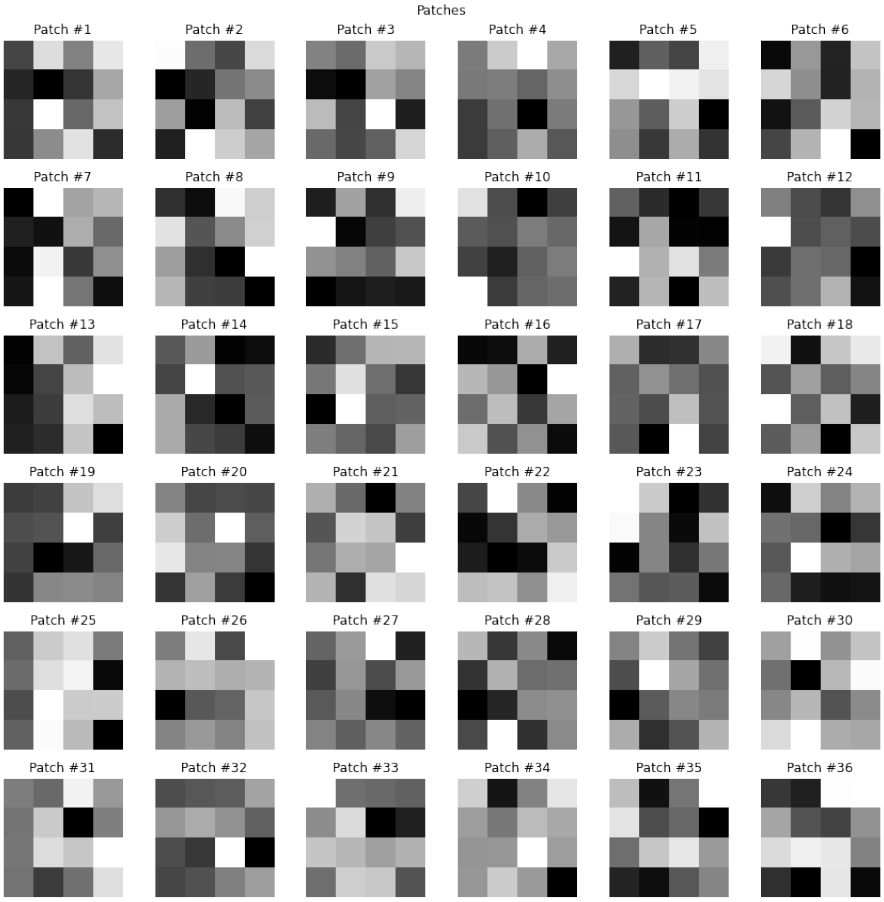

Рис.4.12. Полученный набор патчей.

Теперь проверим гипотезу о том, что распределение числа ненулевых коэффициентов в разреженном представлении для каждого атома действительно удовлетворяет логнормальному распределению. Действительно, по гистограмме распределения интенсивностей, изображенной на рис.4.13, можно заметить, что гипотеза верна. Для каждого разреженного представления вычислялись параметры логнормального распределения, по которым далее на графики добавлялись кривые плотности вероятности, каждая из которая соответствует своему набору параметров.

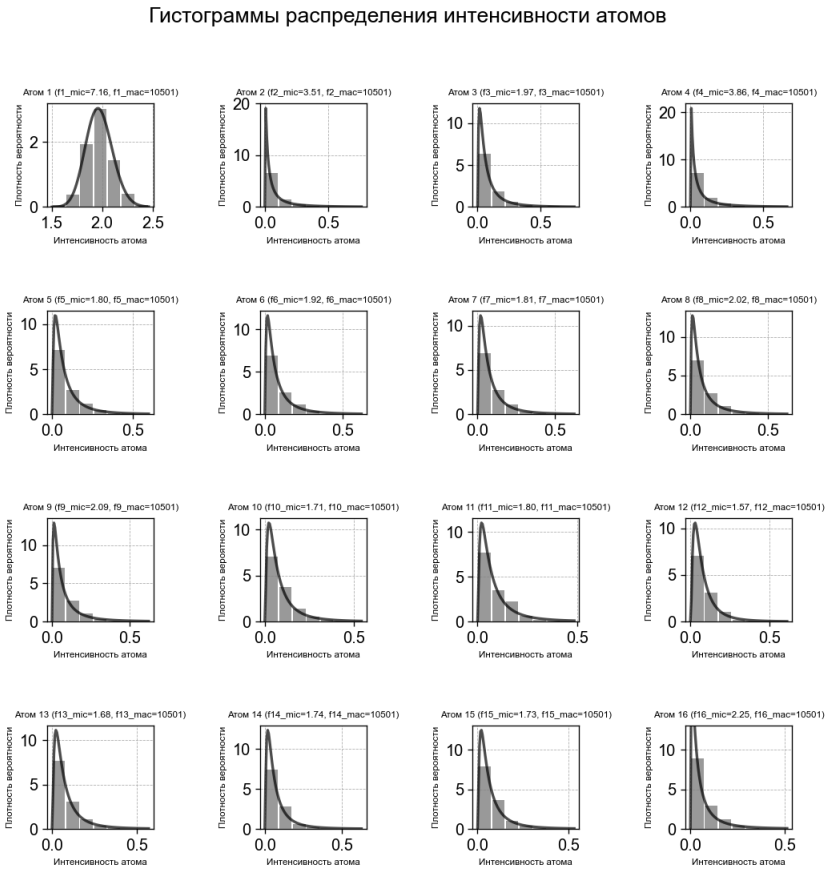

Рис.4.13. Гистограммы распределения интенсивностей.

5) Обучение словаря.

   Далее происходило обучение словаря на полученных патчах. Условием остановки обучения словаря являлось достижение точности равной 0.001. Продемонстрируем для наглядности график обучения словаря на рис.4.14, по нему можно заметить, что заданная точность была достигнута уже на первой итерации обучения, что говорит о высокой эффективности использования K-SVD алгоритма для обучения.

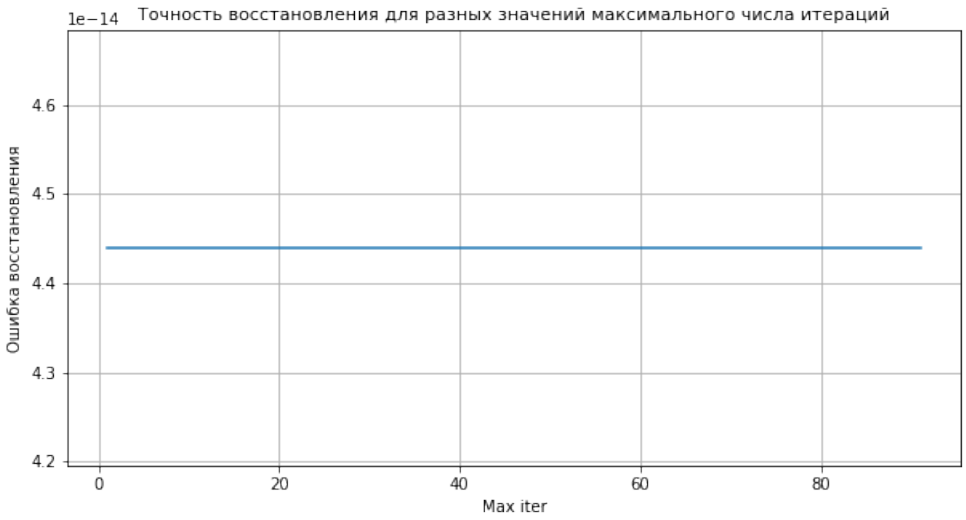

Рис.4.14. Точность восстановления.

Итоговый результат обучения словаря на патчах представлен на рис.4.15.

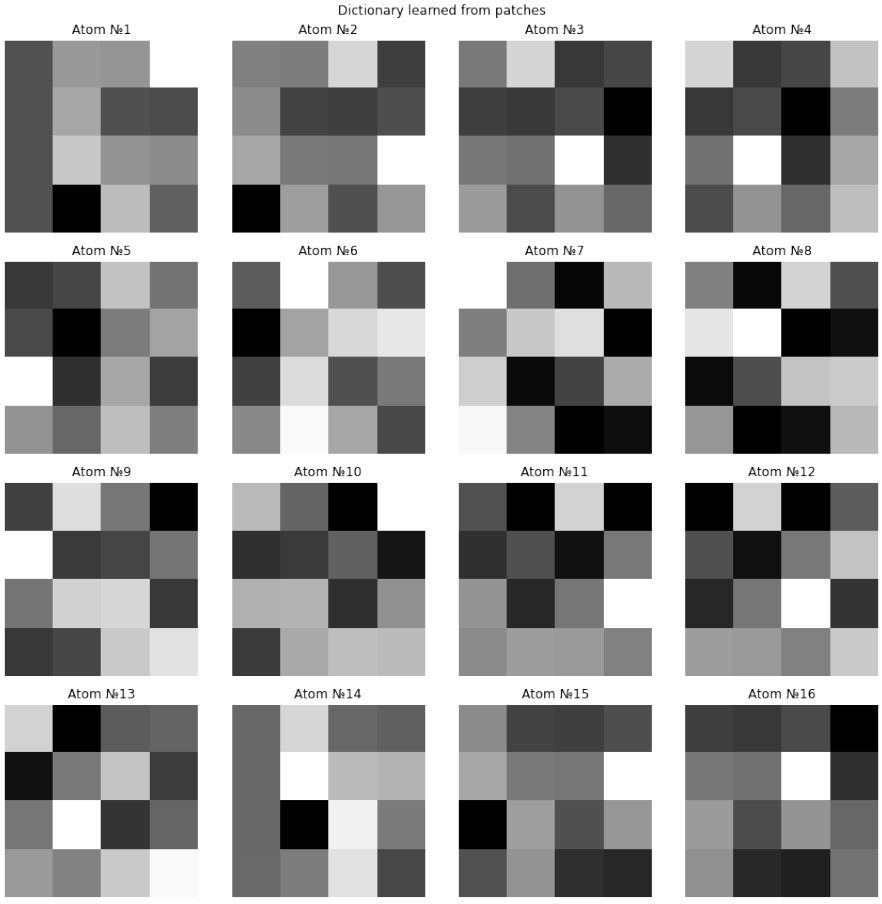

Рис.4.15. Словарь, обученный на патчах.

6) Перечень полученных фич и их корреляция.

Поскольку мы имеем 4 на 4 патч, то микро признаков, так же, как и макро признаков будет по 16 признаков в каждом наборе, что в общем образует 32 фичи.

Теперь проанализируем взаимосвязь между признаками с использованием коэффициента корреляции Пирсона.

Тепловая карта корреляций между признаками представлена на рис.4.16.

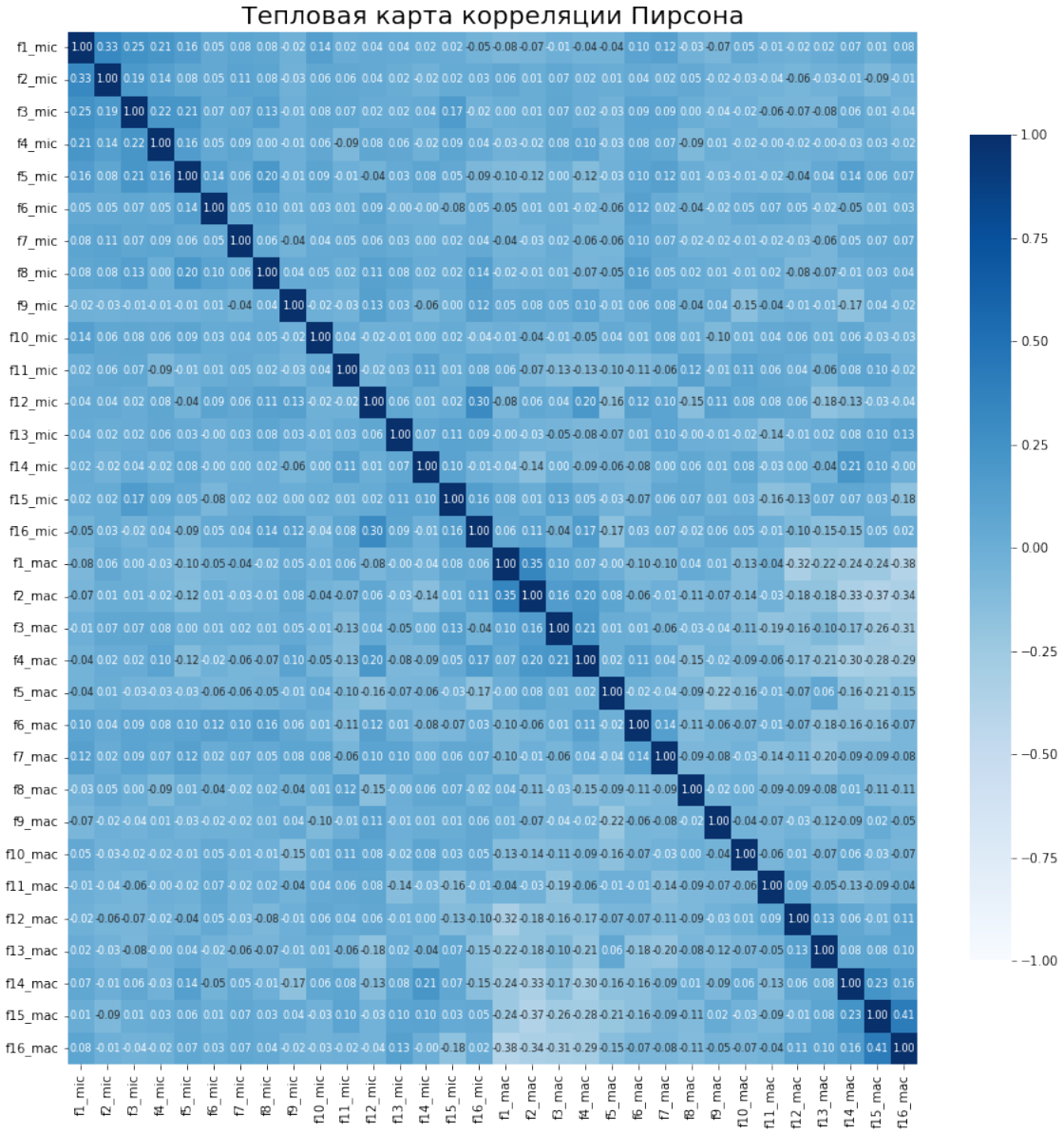

Рис.4.16. Корреляции между признаками.

После проведения анализа корреляции Пирсона и тестирования на статистическую значимость с использованием p-value, результаты были про визуализированы также на тепловой карте (рис.4.17).

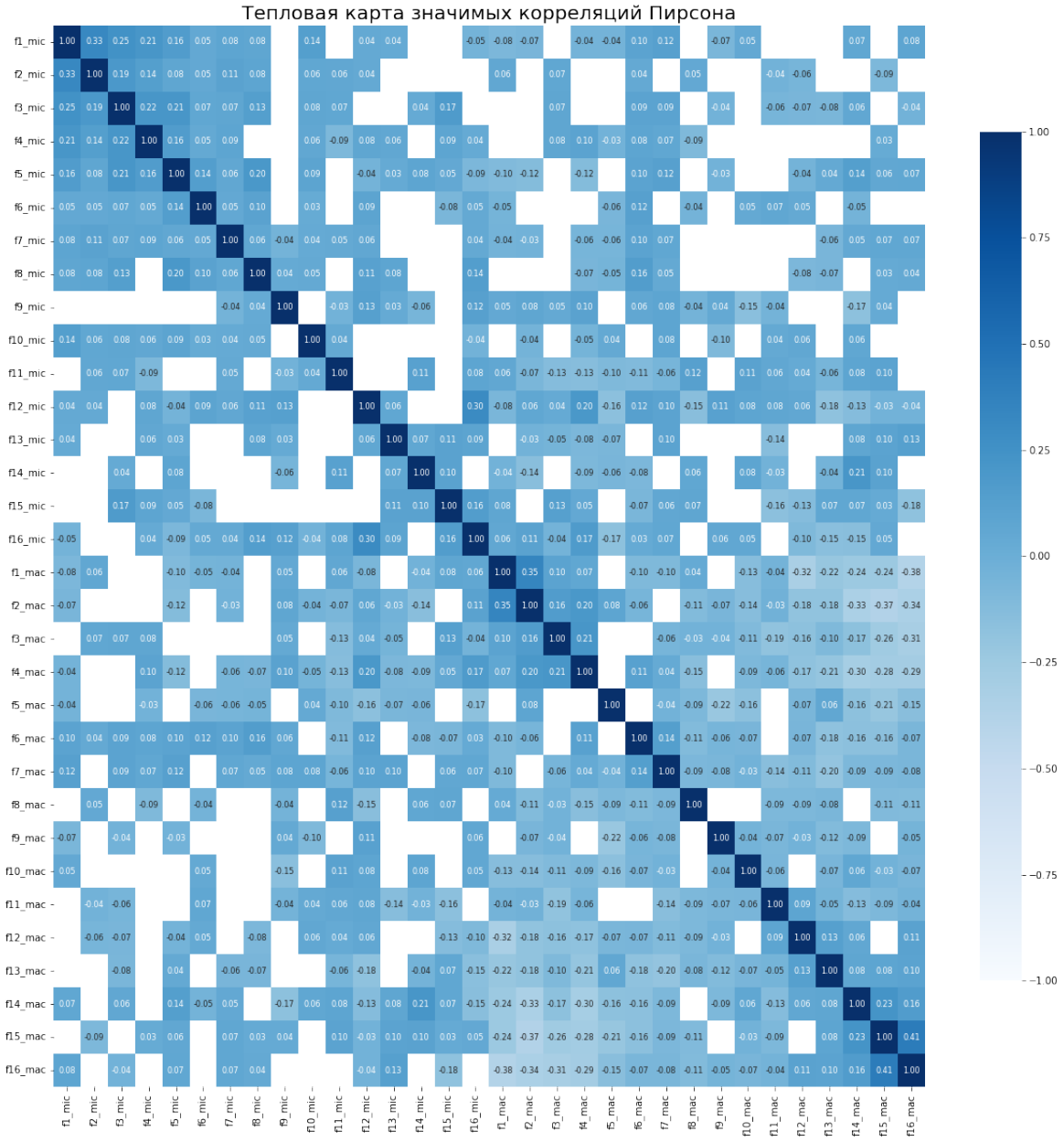

Рис.4.17. Тепловая карта значимых корреляций Пирсона.

В данном исследовании был выбран уровень значимости 0.05, что означает, что можно принять до 5% вероятности ошибки при отвержении нулевой

гипотезы, в данном случае нулевая гипотеза – отсутствие корреляции между двумя признаками.

Данный способ показал наличие взаимосвязей между полученными признаками в наборе данных, которые оказались статистически значимыми. Тем не менее, несмотря на наличие значимых корреляций, было принято решение об использовании всех признаков для дальнейшего анализа. На тепловой карте корреляции Пирсона заметно, что большинство признаков имеют низкие коэффициенты корреляции. Это подтверждает правильность выбора решения включить все признаки в модель, поскольку на данном этапе сложно спрогнозировать, какой конкретно признак внесет существенный вклад в итоговую модель и является наиболее или наименее важным. Вместо исключения признаков на основе p-value на этапе анализа, будут рассматриваться они все в процессе построения регрессионной модели, чтобы не упустить потенциально важные для прогнозирования признаки, исключение признаков на основе только одного критерия может привести к потере важной информации.

7) Получение SVR-модели.

1. K-Fold кросс-валидация и тюнинг гипер-параметров модели.

Для использования кросс-валидации данные были разбиты на 5 фолдов, далее для каждого параметра из сетки заданных параметров происходило обучение модели по всем возможным комбинациям параметров. Наилучшую оценку показала модель со следующими параметрами:

$$C = 1000$$
$$\varepsilon = 1$$
$$\gamma = 1$$

Для данной модели получены следующие показатели метрик на тестовом наборе данных:

Максимальная абсолютная ошибка: 14.9 дней.

Средняя абсолютная ошибка: 4.36 дней.

Средняя квадратическая ошибка: 5.79 дней.

Можем заметить, что точность прогнозирования времени цветения получилась достаточно хорошей, что в целом позволяет применять данную модель в условиях реально поставленных задач.

2. Оценка важности полученных признаков. Следующим этапом ана-

лиза стала оценка важности полученных признаков. Для ее изучения используется подход, основанный на технике перестановки (permutation test). Перестановочный тест является методом статистического анализа, который используется для оценки влияния каждого признака на качество SVR-модели. Основная идея перестановочного теста состоит в том, что наблюдаемые данные переупорядочиваются многократно для создания распределения нулевой гипотезы. Если качество SVR-модели существенно ухудшается после «перемешивания» значений признака, то это указывает на то, что данный признак играет существенную роль для прогностической способности модели. Посмотрим на результаты графика (рис.4.18), который отражает распределение оценок важности признаков.

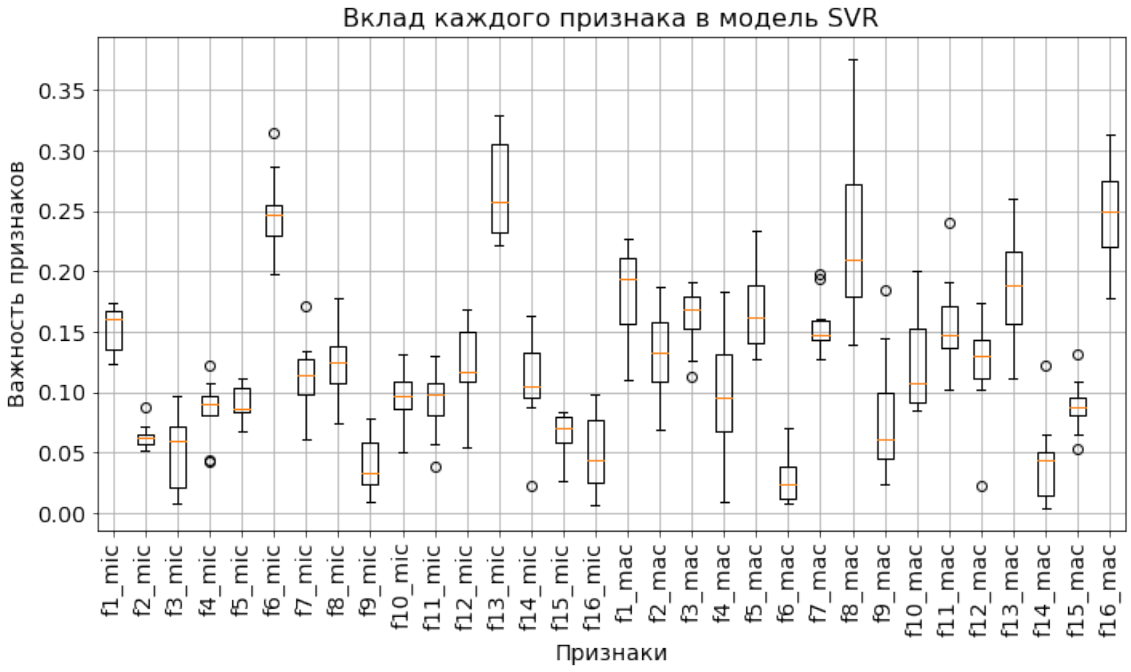

Рис.4.18. Оценка важности вклада признаков в SVR-модель.

Проанализировав его, можно сделать вывод, что изменение порядка значений данных признаков имеет примерно одинаковое влияние на точность модели. Данный факт может означать, что все признаки так или иначе имеют примерно одинаковую важность для модели. Значения влияния каждого признака, полученные в ходе перестановочного теста, на модель получились небольшими, однако это не говорит нам о том, что эти признаки не играют важную роль в прогнозировании, наоборот, каждый из них вносит свой вклад, образуя в совокупности необходимое для обучения и прогнозирования совместное влияние.

Теперь проведем статистический анализ полученных результатов. Применим тест Манна-Уитни для проверки гипотезы о равенстве медиан важности признаков. Нулевая гипотеза состоит в следующем – предполагаем, что важность каждого признака не отличается от среднего значения важности всех признаков, установим уровень значимости равным 0.005. Посмотрим на результаты анализа, представленные на столбчатой диаграмме ниже (рис. 4.19). Для признаков чьи значения получились выше

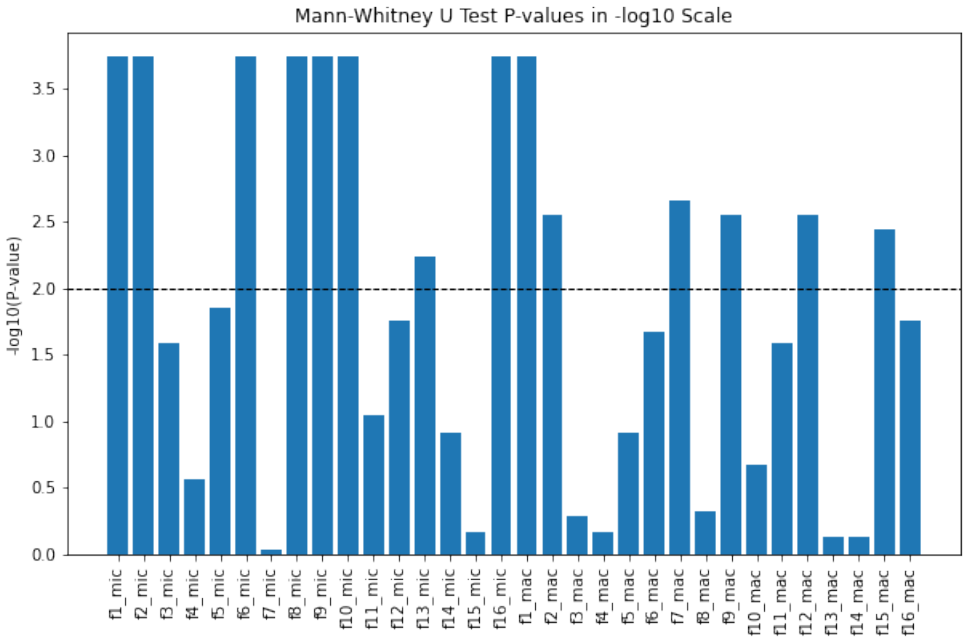

Рис.4.19. Сравнение важности призаков на основе перестановочного теста и теста Манна-Уитни.

линии выбранного нами уровня значимости можем сказать следующее – важность признаков статистически значимо отличается от средней важности всех признаков. Таким образом, можем заметить, что большая часть признаков являются важными для модели, и их вклад не является случайным.

Попробуем исключить те признаки, которые по результатам проведенного статистического теста не вносят существенный вклад в обучение модели, то есть исключаем те фичи, которые лежат ниже пунктирной горизонтальной линии. Посмотрим на вновь полученные результаты метрик модели на тестововом наборе данных:

Максимальная абсолютная ошибка: 13.99 дней.

Средняя абсолютная ошибка: 4.32 дня.

Средняя квадратическая ошибка: 5.45 дня.

Таким образом, можем заметить, что ошибка прогнозирования времени цветения уменьшилась, что подтверждает наши предположения о не сильном влиянии на обучение модели данных признаков, скорее всего, данные признаки вносили некоторый шум в модель при обучении.

3. Реакция модели в условиях имитации климатических изменений.

Проанализируем, как полученная модель реагирует на глобальные изменения, связанные с изменением климатом. Смоделируем климатические изменения, в частности, будем последовательно увеличивать температуру и сокращать выпадение осадков, а затем изучать влияние на модель.

Изменение климата рассматривалось отдельно для каждого местоположения выращивания растений, всего геолокаций в наборе данных три. Продемонстрируем результаты работы модели, представленные на рис.4.20 и рис.4.21.

Несмотря на изменения окружающих условий, модель отражает схожие тенденции изменение фенотипа для всех трех географических локаций, где выращиваются растения. Значение изменений фенотипа колеблются в пределах от 12.5 до 14.5 дней.

Данный анализ показал, что модель в целом адаптируется к климатическим изменениям, но у нее также имеются и некоторые ограничения, связанные с ее постоянной реакцией на изменение климата.

Эта особенность связана с алгоритмом извлечения фич с помощью K-SVD обученного словаря, который фиксирует закономерности обучающих данных, но плохо учитывает новые закономерности, возникающие в результате непрерывных изменений условий окружающей среды.

4. Влияние исходных факторов на модель.

В ходе исследования был проанализирован вклад исходных факторов на модель также с использованием перестановочного теста Манна-Уитни.

Результаты, предоставленные в Приложении 2, показали, что большая часть рассматриваемых исходных факторов, как генетических, так и погодных имеют схожее влияние на модель. Однако удалось выявить одиннадцать ключевых факторов, которые в целом по результатам теста оказали наибольшее влияние на модель.

Генетические факторы определяют базовую "программу"цветения, в то время как погодные факторы могут ускорять или замедлять этот процесс,

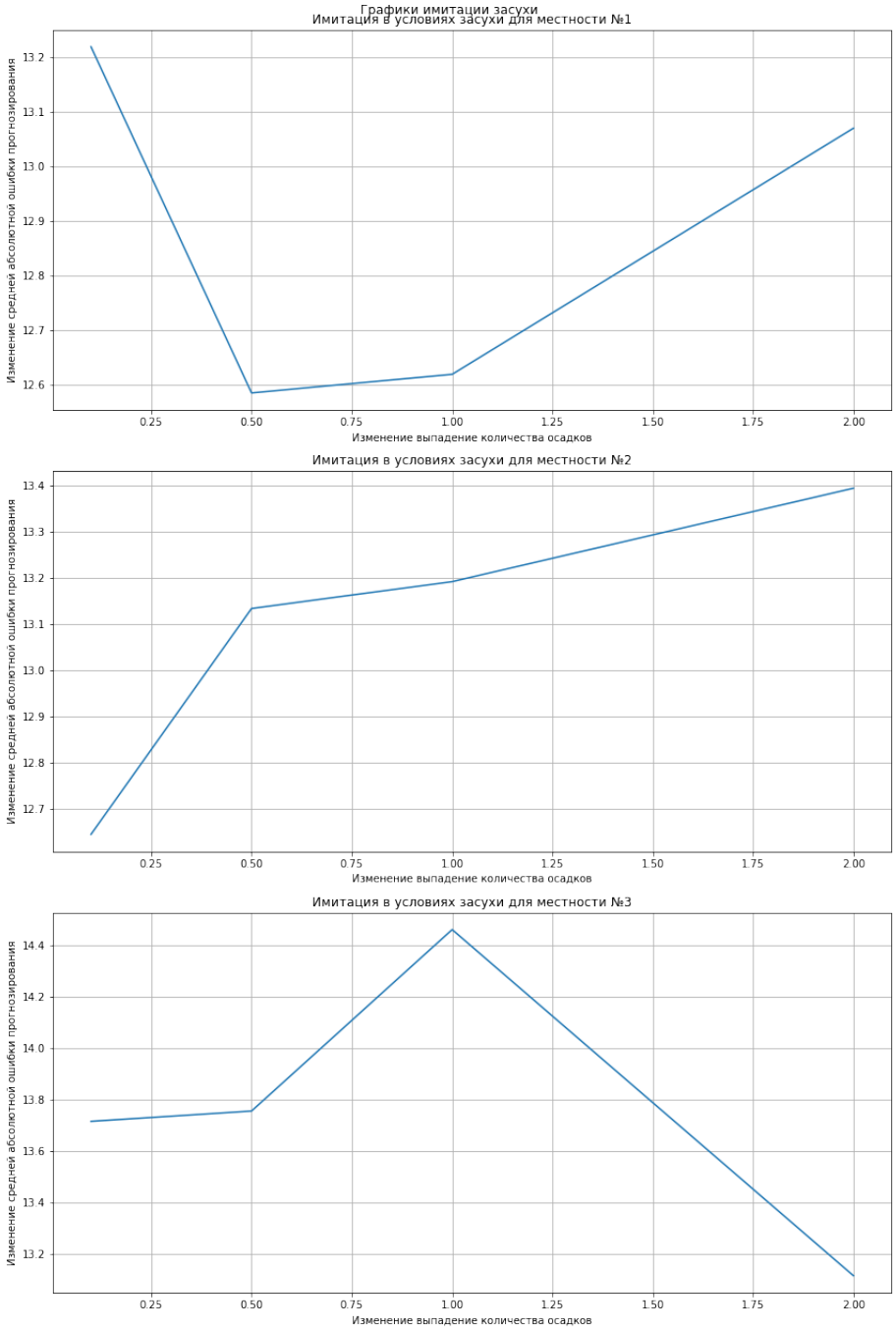

Рис.4.20. Влияние изменения климата на модель, связанные с уменьшением количества осадков.

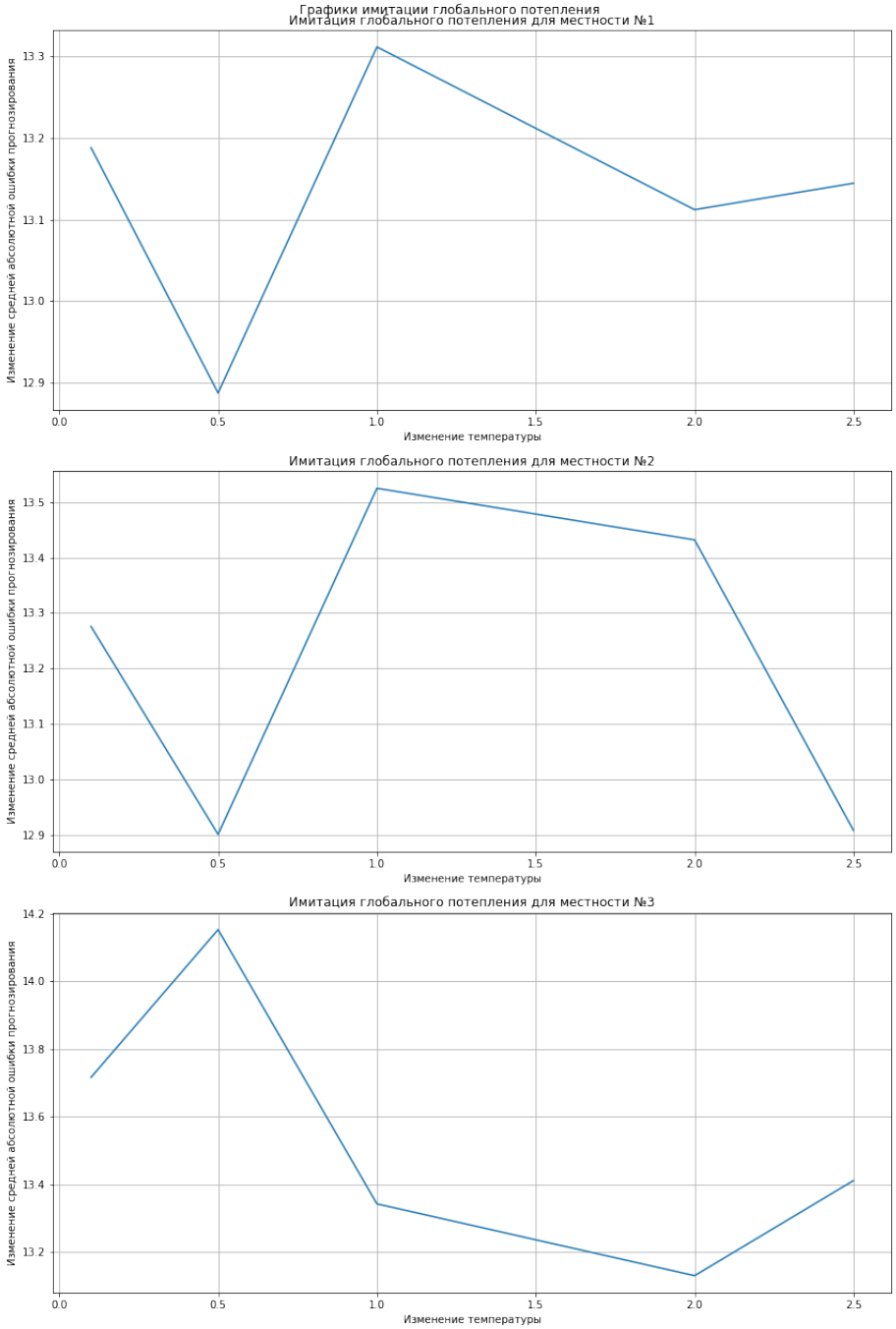

Рис.4.21. Влияние изменения климата на модель, связанные с повышением температуры.

а сочетание этих факторов создает сложную систему, которая определяет время цветения растения.

Несмотря на заметное влияние 11 факторов вклад в модели, важно подчеркнуть, что все рассмотренные факторы тоже являются важными и вносят свой вклад в модель. Вместе с тем, детальный анализ наиболее важных факторов может дать нам больше информации о том, как они взаимодействуют и влияют на фенотип растений, что может помочь нам улучшить качество прогнозирования.

ЗАКЛЮЧЕНИЕ

Таким образом, изучение фенотипов растений, в частности, изучение времени цветения имеет большое значение в современном мире. Учитывая потребности растущего населения планеты, растущий спрос на продукты питания и быстро меняющийся климат, понимание взаимодействия генетических и экологических факторов становится основной задачей.

Предлагаемое исследование направлено на то, чтобы решить данную задачу путем разработки модели прогнозирования фенотипа растений с использованием искусственно сгенерированных изображений. Такая модель представляет новый подход к изучению сельского хозяйства, который может улучшить селекцию растений, их сохранение и наше понимание адаптации и эволюции растений.

Данный алгоритм расположен на стыке технологий и биологии, демонстрируя важность междисциплинарных исследований, что подчеркивает необходимость применения вычислительных методов не только в научных исследованиях, но и в том числе в решении различных прикладных задач, например, биологических.

Разработанная модель основана на оценке качества искусственно сгенерированных изображений, которые кодируют в себе генетические и погодные данные. Для их анализа используется разбиение изображения на патчи (блоки заданного размера) и определение коэффициентов из заранее обученного словаря, который был получен с помощью K-SVD алгоритма обучения словаря, данный алгоритм позволяет найти наилучший словарь, который представляет входной сигнал как линейную комбинацию разреженных представлений. На основе полученных коэффициентов определяется набор признаков, позволяющий с помощью SVR-модели оценить качество изображения.

Хотя разработанная модель хорошо прогнозирует фенотип растений, она имеет некоторые ограничения, связанные с вычислительными затратами, требующимися для вычисления словаря, поэтому в будущем планируется продолжить развитие алгоритма в сторону уменьшения использования памяти для хранения матрицы словаря, это может быть выполнено путем оптимизации разработки искусственно сгенерированных изображений, нахождения оптимальных алгоритмов использования всего набора данных изображений.

ВЫВОДЫ

В работе представлен новый подход к прогнозированию фенотипа растений с использованием разряженного разложения искусственных изображений, которые позволяют закодировать всю предоставленную информацию с помощью использования побитового кодирования данных, затем они подвергаются процессу извлечения признаков посредством эффективного K-SVD алгоритма обучения словаря и разреженного кодирования.

По итогам разработки можно сделать следующие выводы:

- Применение K-SVD алгоритма обучения словаря позволяет добиться необходимой точности на первой итерации.
- С помощью использования методов оптимальной настройки гипер-параметров и процесса K-Fold кросс-валидации удалось построить модель, демонстрирующую высокую точность прогнозирования: средняя ошибка прогнозирования на тестовом наборе данных показала 4.5 дня.
- Модель показала свою устойчивость к смоделированной ситуации глобального потепления, однако отсутствие дальнейшей адаптации модели к последующим изменениям климата может быть связано с тем, что полученный словарь отражает только закономерности в исходных обучающих данных и не фиксирует новые закономерности, вызванные непрерывными изменениями климата. Следовательно, способность модели постоянно адаптироваться к новым шаблонам данных может быть ограничена. Решение проблемы - постоянное обновление словаря с учетом новых данных, отражающих климатические изменения.
- Анализ с применением перестановочного теста Манна-Уитни подтвердил, что большая часть из рассматриваемых исходных факторов вносит одинаковый сопоставимый вклад в модель прогнозирования времени цветения. Однако, в ходе исследования было выявлено одиннадцать ключевых факторов, включая генетические и климатические, которые оказывают наибольшое влияние на модель. Все факторы играют важную роль, но понимание, как именно эти 11 факторов влияют на рост и развитие растений, позволит нам понять механизмы их влияния на фенотип и будет способоствовать более эффективному использованию ресурсов.
- Изучение важности влияния полученных признаков на качество модели позволило найти те избыточные предикторы, отрицательно влияющие

на обощающее свойство модели. С помощью статистического анализа теста Манна-Уитни удалось их исключить и посмотреть на новое качество модели, которое уменьшило максимальную ошибку прогнозирования на 1 день, в итоге максимальная ошибка прогнозирования времени цветения стала - 13.99 дня, средняя абсолютная ошибка так же уменьшилась, но на десятые доли и составляет - 4.32 дня.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Phenotypic plasticity in plant height shaped by interaction between genetic loci and diurnal temperature range / Q. Mu [et al.] // New Phytologist. — 2021. — Vol. 233, issue 4. — P. 1768–1779. — DOI 10.1111/nph.17904.

2. A multiscale analysis of early flower development in Arabidopsis provides an integrated view of molecular regulation and growth control / Y. Refahi [et al.] // Developmental Cell. — 2021. — Vol. 56, issue 4. — P. 540–556. — DOI 10.1016/j.devcel.2021.01.019.

3. *Cho L.-H.*, *Yoon J.*, *An G.* The control of flowering time by environmental factors // The Plant Journal. — 2017. — Vol. 90, issue 4. — P. 708–719. — DOI 10.1111/tpj.13461.

4. What is vernalization and how do plants remember winter? — URL: https://www.jic.ac.uk/blog/what-is-vernalization-and-how-do-plants-remember-winter (дата обращения: 30.05.2023).

5. Photoperiodism. — URL: https://www.sciencefacts.net/photoperiodism.html (дата обращения: 30.05.2023).

6. *Matesanz S.*, *Gianoli E.*, *Valladares F.* Global change and the evolution of phenotypic plasticity in plants // Annals of the New York Academy of Sciences. — 2010. — Vol. 1206, issue 1. — P. 35–55. — DOI 10.1111/j.1749-6632.2010.05704.x.

7. *Prevéy J.-S.* Climate Change: Flowering Time May Be Shifting in Surprising Ways // Current Biology. — 2020. — Vol. 30, issue 3. — P. 112–114. — DOI 10.1016/j.cub.2019.12.009.

8. Sustainable intensification with irrigation raises farm profit despite climate emergency / A. Muleke [et al.] // Plants, People, Planet. — 2023. — Vol. 5, issue 3. — P. 368–385. — DOI 10.1002/ppp3.10354.

9. Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features / W. Xue [et al.] // IEEE Trans Image Process. — 2014. — Vol. 23, no. 11. — P. 4850–4862. — DOI 10.1109/TIP.2014.2355716.

10. Blind Quality Assessment of Screen Content Images Via Macro-Micro Modeling of Tensor Domain Dictionary / Y. Bai [et al.] // IEEE Transactions on Multimedia. — 2021. — Vol. 23. — P. 4259–4271. — DOI 10.1109/tmm.2020.3039382.

11. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index / W. Xue [et al.] // IEEE Trans Image Process. — 2014. — Vol. 23, issue 2. — P. 684–695. — DOI 10.1109/TIP.2013.2293423.

12. Gradient Direction for Screen Content Image Quality Assessment / N. Zhangkai [et al.] // IEEE Signal Processing Letters. — 2016. — Vol. 23, no. 10. — P. 1394–1398. — DOI 10.1109/LSP.2016.2599294.

13. ESIM: Edge Similarity for Screen Content Image Quality Assessment / N. Zhangkai [et al.] // IEEE Transactions on Image Processing. — 2017. — Vol. 26, no. 10. — P. 4818–4831. — DOI 10.1109/TIP.2017.2718185.

14. Screen Content Image Quality Assessment Using Multi-scale Difference of Gaussian / F. Ying [et al.] // IEEE Transactions on Circuits and Systems for Video Technology. — 2018. — Vol. 28, no. 9. — P. 2428–2432. — DOI 10.1109/TCSVT.2018.2854176.

15. Objective Quality Assessment of Screen Content Images by Uncertainty Weighting / Y. Fang [et al.] // IEEE Transactions on Image Processing. — 2017. — Vol. 26, issue 4. — P. 2016–2017. — DOI 10.1109/TIP.2017.2669840.

16. *Zhang Y.*, *Chandler D.*, *Mou X.* Quality Assessment of Screen Content Images via Convolutional-Neural-Network-Based Synthetic/Natural Segmentation // IEEE Transactions on Image Processing. — 2018. — Vol. 27, no. 10. — P. 5113–5128. — DOI 10.1109/TIP.2018.2851390.

17. Image Sharpness Assessment by Sparse Representation / L. Li [et al.] // IEEE Transactions on Multimedia. — 2016. — Vol. 18, no. 6. — P. 1085–1097. — DOI 10.1109/TMM.2016.2545398.

18. *Lee D.*, *Plataniotis K.* Toward a No-Reference Image Quality Assessment Using Statistics of Perceptual Color Descriptors // IEEE Transactions on Image Processing. — 2016. — Vol. 25, no. 8. — P. 3875–3889. — DOI 10.1109/TIP.2016.2579308.

19. *Mittal A.*, *Moorthy A.*, *Bovik A.* No-Reference Image Quality Assessment in the Spatial Domain // IEEE Transactions on Image Processing. — 2012. — Vol. 21, no. 12. — P. 4695–4708. — DOI 10.1109/TIP.2012.2214050.

20. No reference quality evaluation for screen content images considering texture feature based on sparse representation / J. Yang [et al.] // Signal Processing. — 2018. — Vol. 153. — P. 336–347. — DOI 10.1016/j.sigpro.2018.07.006.

21. A Gabor Feature-Based Quality Assessment Model for Screen Content Images / Z. Ni [et al.] // IEEE Transactions on Image Processing. — 2018. — Vol. 27, no. 9. — P. 4516–4528. — DOI 10.1109/TIP.2018.2839890.

22. Screen Content Image Quality Assessment With Edge Features in Gradient Domain / R. Wang [et al.] // IEEE Access. — 2019. — Vol. 7. — P. 5285–5295. — DOI 10.1109/ACCESS.2018.2889992.

23. *Sharifi K.*, *Leon-Garcia A.* Estimation of Shape Parameter for Generalized Gaussian Distributions in Subband Decompositions of Video // IEEE Transactions on Circuits and Systems for Video Technology. — 1995. — Vol. 5, no. 1. — P. 52–56. — DOI 10.1109/76.350779.

24. *Guo H.*, *Ma K.-K.*, *Zeng H.* A Log-Gabor Feature-Based Quality Assessment Model for Screen Content Images // 2019 IEEE International Conference on Image Processing (ICIP). — IEEE. Taipei, Taiwan, 2019. — P. 4499–4503. — DOI 10.1109/ICIP.2019.8803491.

25. Visual Information Measurement with Quality Assessment / J. Wu [et al.] // 2016 Visual Communications and Image Processing (VCIP). — IEEE. Chengdu, China, 2016. — P. 1–4. — DOI 10.1109/VCIP.2016.7805469.

26. Local and Global Feature Learning for Blind Quality Evaluation of Screen Content and Natural Scene Images / W. Zhou [et al.] // IEEE Transactions on Image Processing. — 2018. — Vol. 27, no. 5. — P. 2086–2095. — DOI 10.1109/TIP.2018.2794207.

27. Blind Image Quality Assessment of Screen Content Images via Fisher Vector Coding / Y. Bai [et al.] // IEEE Access. — 2022. — Vol. 10. — P. 13174–13181. — DOI 10.1109/ACCESS.2022.3141914.

28. Image Quality Assessment: From Error Visibility to Structural Similarity / Z. Wang [et al.] // IEEE Transactions on Image Processing. — 2004. — Vol. 13, no. 4. — P. 600–612. — DOI 10.1109/TIP.2003.819861.

29. *Aharon M.*, *Elad M.*, *Bruckstein A.* K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation // IEEE Transactions on Signal Processing. — 2006. — Vol. 54, no. 11. — P. 4311–4322. — DOI 10.1109/TSP.2006.881199.

30. Tuning the hyper-parameters of an estimator. — URL: https://scikit-learn.org/stable/modules/grid_search.html (дата обращения: 05.05.2023).

Приложение 1

Блок-схема алгоритма оценки качества изображений

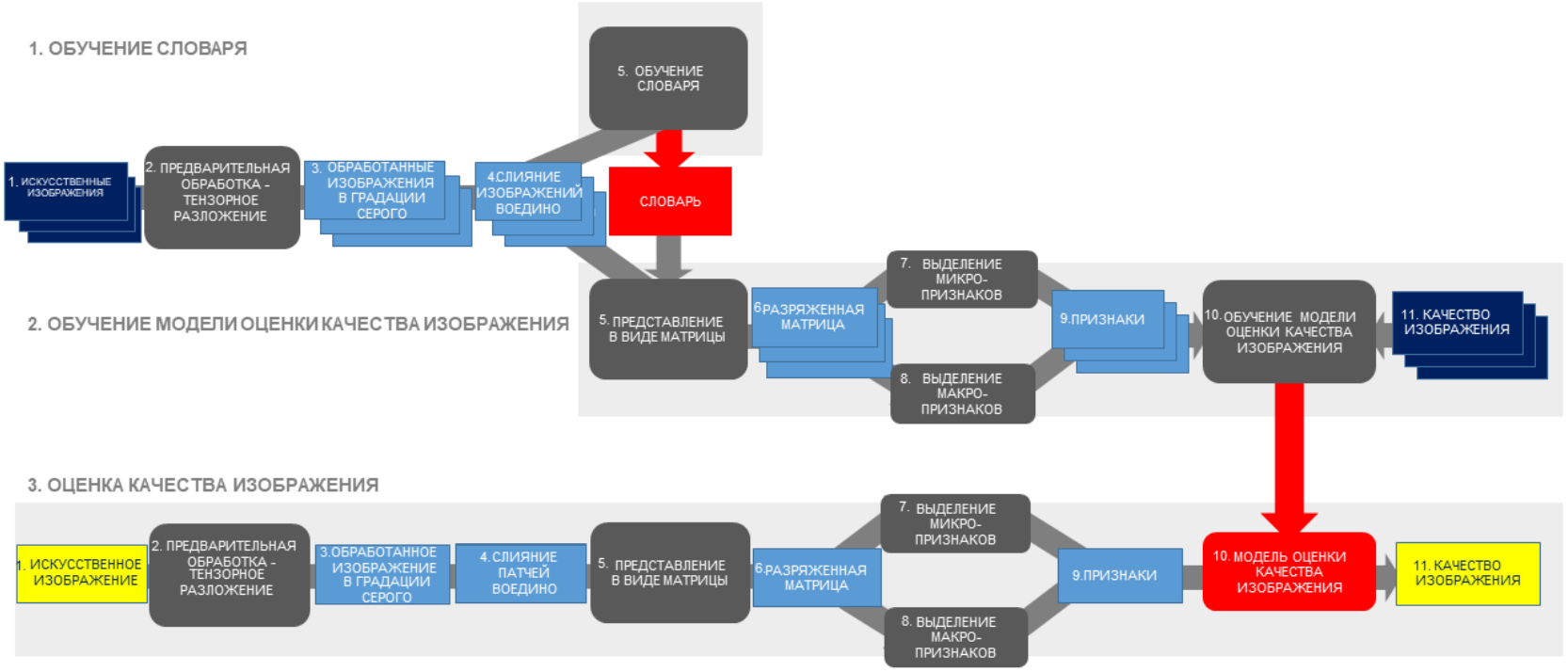

Рис.П1.1. Структура алгоритма оценки качества изображений.

Приложение 2

Влияние исходных факторов на модель прогнозирования

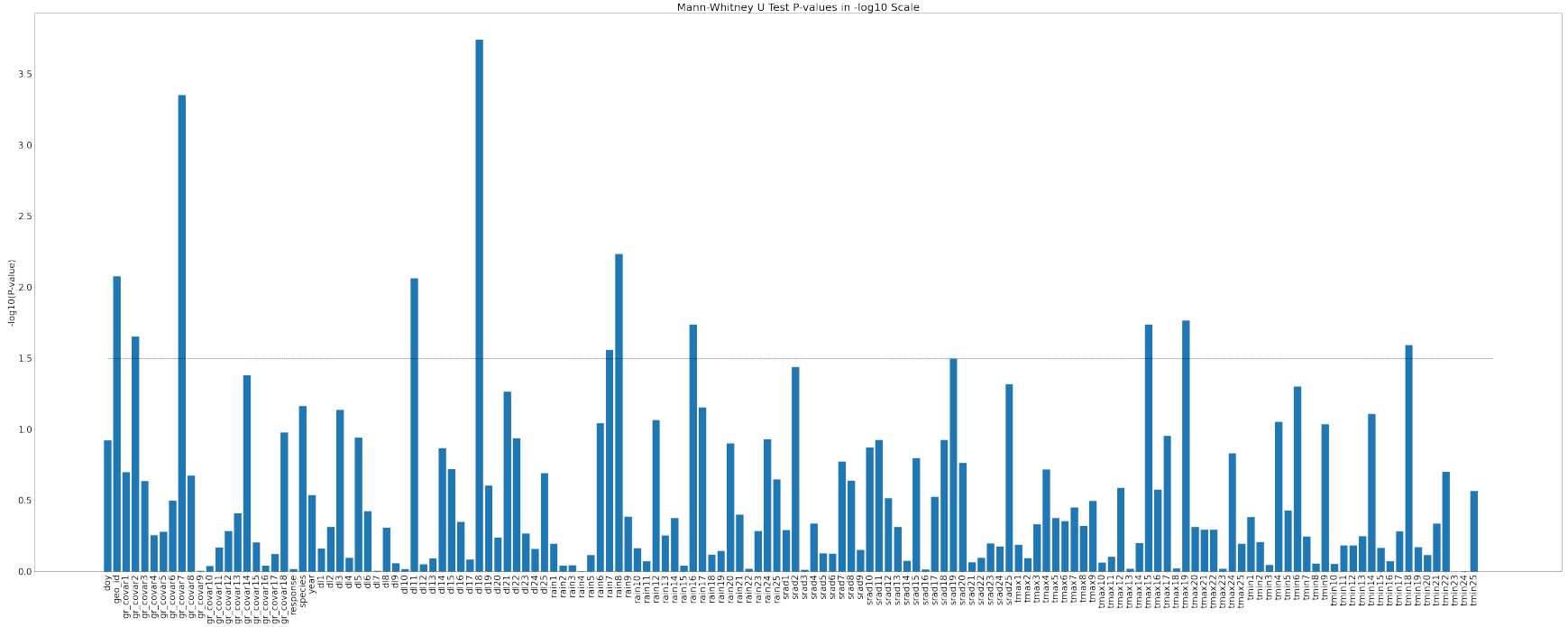

Рис.П2.1. Результаты перестановочного теста Манна-Уитни