УДК 004.724

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРЕДСКАЗАНИЯ ХИМИЧЕСКОГО СОСТАВА МАТЕРИАЛОВ

В. П. Корячко, д.т.н., профессор, заведующий кафедрой САПР ВС РГРТУ, Рязань, Россия;
orcid.org/0000-0000-0000-000X, e-mail: koryachko.v.p@rsreu.ru
С. Д. Викулин, аспирант РГРТУ, Рязань, Россия;
orcid.org/0009-0002-9932-1113, e-mail: vikulin97@gmail.ru
А. В. Волков, специалист, Москва, Россия;
orcid.org/0009-0008-1162-3816, e-mail: vic-volk@yandex.ru

Рассматривается задача предсказания химического состава материалов на основе их физических свойств с использованием методов машинного обучения, включая многоклассовую регрессию. ***Целью данной работы*** *является изучение взаимосвязей между физико-химическими характеристиками материалов и их химическим составом для разработки интеллектуальной системы, поддерживающей процессы проектирования новых материалов. В исследовании использованы методы линейной регрессии, дерева решений и нейронных сетей для предсказания химического состава различных материалов. Результаты показали, что предложенные методы позволяют с высокой точностью предсказать состав материалов, что открывает перспективы для оптимизации процессов разработки и улучшения характеристик материалов.*

Ключевые слова: *машинное обучение, многоклассовая регрессия, химический состав, физические свойства материалов, нейронные сети, линейная регрессия, дерево решений, материаловедение.*

DOI: 10.21667/1995-4565-2025-91-50-63

Введение

Развитие передовых материалов становится всё более значимым фактором в таких отраслях, как автомобилестроение, авиакосмическая промышленность, медицина и энергетика. Современные технологии предъявляют повышенные требования к характеристикам материалов, что стимулирует учёных к созданию новых методов их исследования и разработки. Передовые материалы должны обладать уникальными свойствами, включая улучшенные механические, термические и химические характеристики, чтобы соответствовать задачам высокотехнологичных отраслей. Это требует инструментов, способных проводить микроструктурный анализ, прогнозировать свойства и улучшать производственные процессы с высокой скоростью и точностью [1].

Искусственный интеллект (ИИ), особенно нейросетевые модели, демонстрирует высокий потенциал в этой области, поскольку они способны быстро оценивать свойства материалов, предсказывать их параметры и открывать новые закономерности [2]. Благодаря возможностям учета локальных и глобальных взаимодействий в структурах материалов, нейросети успешно применяются для прогнозирования свойств молекул, материалов и их реакционной способности [3]. Эти достижения открывают новые возможности для проектирования материалов с заданными характеристиками и значительно ускоряют процесс исследований.

Свойства материалов, такие как механическая прочность, модуль Юнга, теплопроводность и плотность, играют ключевую роль в оценке их применимости в различных условиях [4]. Например, понимание корреляции между упругими и термическими свойствами позволяет прогнозировать механическую стабильность и износостойкость [5]. Кроме того такие параметры, как радиусы электронных орбиталей, помогают прогнозировать пластичность материалов, что значительно упрощает процесс их разработки [6].

Одной из наиболее сложных задач материаловедения является предсказание химического состава материала на основе известных свойств. Решение этой задачи важно для создания новых сплавов, композитов и оптимизации технологических процессов. Методы искусственного интеллекта в сочетании с экспериментальными данными уже доказали свою эффективность в этой области. Например, искусственный интеллект позволяет создавать модели для прогнозирования химического состава сталей для минимизации затрат на испытание этих сталей [7]. Также системы рекомендаций, основанные на машинном обучении, могут предоставлять инженерам данные о возможных сочетаниях свойств и химического состава, минимизируя затраты на эксперименты и ускоряя цикл разработки материалов [8].

Таким образом, использование методов искусственного интеллекта становится стандартом в современных исследованиях материалов. Эти подходы позволяют не только анализировать существующие закономерности, но и выявлять новые взаимосвязи, которые были недоступны традиционным методам. В частности, прогнозирование химического состава материалов с использованием искусственных нейронных сетей открывает перспективы создания материалов с уникальными свойствами, что становится важным инструментом для решения актуальных научных и технологических задач.

Целью данной работы является изучение взаимосвязей между основными свойствами материалов и их химическим составом для разработки интеллектуальной системы, способной поддерживать процессы проектирования новых материалов.

Постановка задачи

Необходимо исследовать и сравнить методы машинного обучения и нейронных сетей для предсказания химического состава материалов на основе их физических свойств с использованием подхода многоклассовой регрессии. В качестве объектов исследования выбраны металлические материалы и сплавы, содержащие химические элементы, такие как алюминий, углерод, хром, медь, железо, магний, титан, цинк и другие. Для описания материалов были отобраны следующие ключевые физические свойства: плотность, модуль упругости, теплопроводность, электрическое сопротивление и удельная теплоемкость. Каждое из этих свойств является важным для характеристики материалов и их применения в различных отраслях, таких как машиностроение, авиастроение, металлургия и др.

Задача состоит в том, чтобы на основе данных о физических свойствах материалов точно предсказать пропорции химических элементов в их составе. Для решения задачи будут использованы различные алгоритмы машинного обучения, включая линейную регрессию, дерево решений и нейронные сети. Дополнительно будет исследован процесс оптимизации гиперпараметров моделей, так как правильная настройка гиперпараметров может существенно повысить точность предсказаний.

Теоретические исследования

Мультиномииальная (многоклассовая) регрессия является простым расширением стандартной логистической регрессии для решения задач многоклассовой классификации. В контексте данной задачи выходной вектор модифицируется в вектор значений, являющийся пропорциями химических элементов, который соответствует вероятностному распределению при сумме всех значений равных единице [9].

Таким образом имеется матрица признаков $X \in \mathbb{R}^{n \times m}$ и матрица пропорций химических элементов $Y \in \mathbb{R}^{n \times k}$, где n – количество элементов в выборке (материалов), m – количество физических свойств (признаков), k – количество химических элементов (выходы). Для корректного обучения моделей применим Z-преобразования для каждого элемента x_i матрицы признаков. Нормализированное значение x_i' вычисляется следующим образом:

$$x_i' = \frac{x_i - \mu}{\sigma}, \tag{1}$$

где $\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$ – среднее значение признака x_i; $\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2}$ – стандартное отклонение признака.

Для моделирования пропорций химических элементов используем софтмакс-функцию (softmax), которая будет гарантировать, что выходные значения находятся в диапазоне [0,1] и их сумма равна 1:

$$\widehat{y}_{ij} = \frac{\exp(x_i'\beta_j)}{\sum_{i=1}^{k}\exp(x_i'\beta_j)}, \quad (2)$$

где $\widehat{y}_{ij}$ – предсказанная доля химического элемента j для материала i; $x_i'\beta_j$ – линейное преобразование признаков x_i' с использованием весового вектора β_j; β_j – вектор параметров для химического элемента j; $\sum_{l=1}^{k}\exp(x_i'\beta_j)$ – нормирующий член, который гарантирует, что сумма всех выходных значений равна 1.

Для оценки качества модели используется функция потерь, основанная на отрицательной логарифмической правдоподобности, так как мультиномиальная регрессия рассматривает задачу как вероятностную:

$$J(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log(\widehat{y}_{ij}). \quad (3)$$

Таким образом, из (2) и (3) следует, что задача мультиномиальной регрессии в контексте предсказания химического состава материалов сводится к следующей оптимизационной задаче:

$$\min_{\beta} J(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log\left(\frac{\exp(x_i'\beta_j)}{\sum_{l=1}^{k}\exp(x_i'\beta_j)}\right). \quad (4)$$

Представленный математический аппарат полностью описывает задачу многоклассовой регрессии и её реализацию в данном исследовании.

В качестве методов решения данной задачи будут рассматриваться классическая линейная регрессия с модификацией метода наименьших квадратов регуляризирующим штрафом (RidgeRegression), метод деревьев решения (RandomForest) и собственная реализация нейронной сети (MLP).

Целевая функция потерь для RidgeRegression записывается в следующем виде:

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \lambda\sum_{j=1}^{k}\beta_j^2, \quad (5)$$

где λ – коэффициент регуляризации. Так как классическая линейная регрессия минимизирует сумму квадратов ошибок, то задача сводится из (5) к оптимизации функции:

$$\beta = \arg\min_{\beta}\frac{1}{n}\|Y - X\beta\|^2 + \lambda\|\beta\|^2. \quad (6)$$

Дерево решений представляет собой последовательный процесс разбиения данных на подмножества на основе значений признаков. Каждое разбиение (ветвь) принимает решение, уменьшая сложность задачи и приближая нас к целевому значению. На каждом узле дерева решений выбирается признак, который разделяет данные на две (или более) группы так, чтобы минимизировать определённый критерий разбиения. Этот критерий оценивает качество разделения данных. Для задачи многоклассовой регрессии используется мера ошибки E, ко-

торая вычисляется для каждого узла и позволяет определить, насколько хорошо данные разделены:

$$E = \sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \hat{y}_{ij})^2 . \quad (7)$$

Алгоритм RandomForest строит множество деревьев решений и объединяет их результаты для улучшения качества моделей. Для задачи многоклассовой регрессии выходное значение рассчитывается как среднее предсказание всех деревьев в ансамбле:

$$\hat{y}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_{ij}^{t} , \quad (8)$$

где T – количество деревьев в ансамбле, $\hat{y}_{ij}^{t}$ – предсказание t-го дерева для пропорции химического элемента j объекта i.

Для улучшения обобщающей способности модели в алгоритме RandomForest используется ключевая особенность – случайное подмножество признаков при разбиении в каждом узле. В отличие от традиционного дерева решений, где для каждого разбиения используется весь набор признаков, в RandomForest на каждом шаге выбирается случайное подмножество признаков, что уменьшает корреляцию между деревьями и способствует лучшему обобщению модели. Этот подход позволяет создать более разнообразную коллекцию деревьев, что улучшает устойчивость модели к переобучению и делает ее более точной при работе с новыми данными [10].

Последним методом решения задачи является нейронная сеть. Процесс обучения осуществляется в рамках задачи обучения с учителем, где сеть минимизирует функцию потерь, оценивающую расхождение между реальными и предсказанными значениями. Целевая функция потерь L выбрана как среднеквадратичная ошибка:

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 . \quad (9)$$

В качестве входного используется слой размерности вектора Xc функций активаций ReLU;

$$f(x) = \max(0, x) . \quad (10)$$

Каждый скрытый слой 1 описывается как:

$$h^l = f(W^l \cdot h^{l-1} + b^l) , \quad (11)$$

где W^l – матрица весов, b^l – матрица смещения, f – функция активации (10). Выходной слой использует вышеупомянутую функцию softmax (2). Регуляризация достигается за счет введения слоев исключения и нормализации. Для оптимизации используется популярный и эффективный алгоритм Adam за счет устойчивости к шуму и способности быстро сходиться [11]. Данный алгоритм использует стохастический градиентный спуск с настройкой скорости обучения η :

$$W \leftarrow W - \eta \cdot \frac{\partial L}{\partial W}, \qquad b \leftarrow b - \eta \cdot \frac{\partial L}{\partial b} . \quad (12)$$

В качестве функции потерь была выбрана функция категориальной кросс-энтропии, основанная на отрицательной логарифмической правдоподобности, что было указано в (3). Она широко применяется для многоклассовой классификации и регрессии [12]. Архитектура схематично представлена на рисунке 1.

Для оценки эффективности алгоритмов были использованы следующие оценки качества.

– Среднеквадратичная ошибка (MSE). Равна функции потерь L(9). Оценивает точность предсказаний.

– Средняя абсолютная ошибка. Оценивает устойчивость к выбросам.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - \hat{y}_i | \quad (13)$$

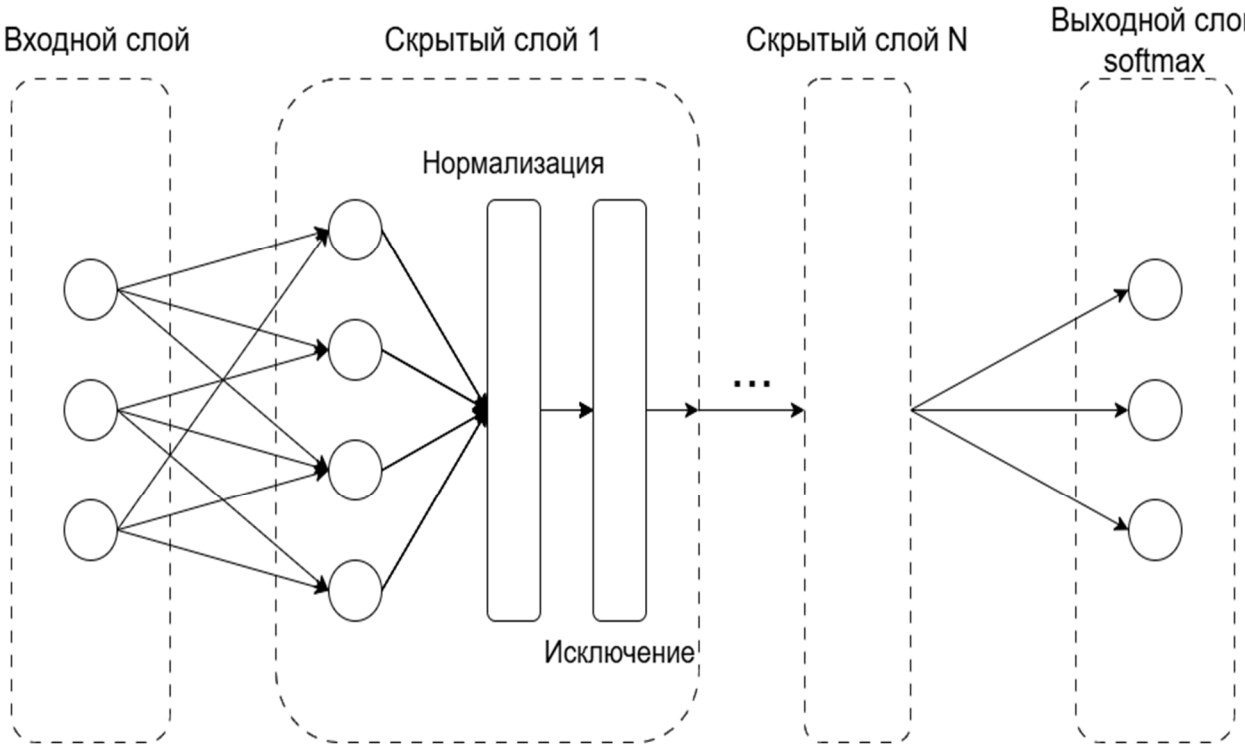

Рисунок 1 –Схема архитектуры нейронной сети для многоклассовой регрессии
Figure 1 – Neural network architecture diagram for multiclass regression

– Коэффициент детерминации. Показывает долю дисперсии истинных значений, объясняемых моделью.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (14)$$

– где $\bar{y}_i$ – среднее значение истинных значений

В качестве гиперпараметрической оптимизации нейронной сети был выбран метод GridSearch. Этот метод создает сетку всех возможных комбинаций гиперпараметров и обучает модель для каждой комбинации. Затем оценивается качество модели для каждой конфигурации. Так как пространство поиска невелико, то имеется возможность протестировать все возможные комбинации сети, такие как количество нейронов и слоев, коэффициент скорости обучения и коэффициент исключения [13].

Экспериментальные исследования

Реальные данные часто бывают неполными, зашумленными и противоречивыми. Для исходной коллекции материалов была проведена предварительная очистка с целью обеспечения корректности и достоверности дальнейшего анализа.

В первую очередь было проведено удаление дубликатов, что позволило повысить качество данных и предотвратить возможные ошибки, связанные с избыточными значениями. Этот шаг особенно важен, если данные поступают из разных источников или были собраны с использованием разных методов, что может привести к повторяющимся записям [14].

Во-вторых, была проведена очистка от нулевых значений в исходных признаках. Кроме того, был применен подход с удалением столбцов, содержащих значительное количество пропусков. Это позволяет исключить избыточные столбцы, которые не несут полезной информации, и сохранить наиболее информативные данные. В частности, элементы материалов, которые встречаются реже, чем в 5 % случаев, были объединены в общий столбец [15].

Это решение способствует улучшению качества анализа, так как столбцы, в которых слишком много пропусков, будут мешать качественному анализу.

Далее для выявления выбросов был использован метод межквартильного размаха (IQR) [16]. Для каждого из выбранных признаков, таких как плотность, теплопроводность, модуль упругости и удельная теплоемкость, был рассчитан IQR, и выбросы были определены как значения, выходящие за пределы пятиквартильного размаха. Выбросы были визуализированы с помощью тепловой карты на рисунке 2, где применялись два цвета: желтый для значений, соответствующих выбросам, и синий для остальных значений.

На рисунке 2 можно наблюдать незначительное количество выбросов по таким параметрам, как плотность и электрическое сопротивление. Все выявленные выбросы были удалены из исходной коллекции для обеспечения корректности анализа.

Следующим этапом стал анализ корреляции между входными параметрами. Для этого был вычислен коэффициент корреляции Пирсона для всех признаков. Результат был визуализирован с помощью тепловой карты на рисунке 3, где значения коэффициентов корреляции были отражены различной насыщенностью цвета.

На тепловой карте корреляции видно отсутствие сильных зависимостей между признаками, что является важным условием для построения эффективной модели. Отсутствие значительной корреляции гарантирует устойчивость алгоритма и исключает необходимость удаления избыточных критериев.

После этого была выполнена нормализация данных для приведения всех параметров к единому масштабу. Это позволило улучшить сходимость алгоритмов и повысить точность модели.

Для анализа распределения химического состава материалов в коллекции были вычислены суммарные значения для каждого химического элемента. Сумма по каждому компоненту отображает общий вклад этого элемента в состав всех образцов. Эти компоненты визуализированы в виде столбчатой диаграммы, где на оси X отображаются химические элементы, а на оси Y — их суммарное значение. Такая визуализация (рисунок 4) позволяет легко оценить относительное содержание каждого элемента и является важным шагом для дальнейшего анализа состава материалов.

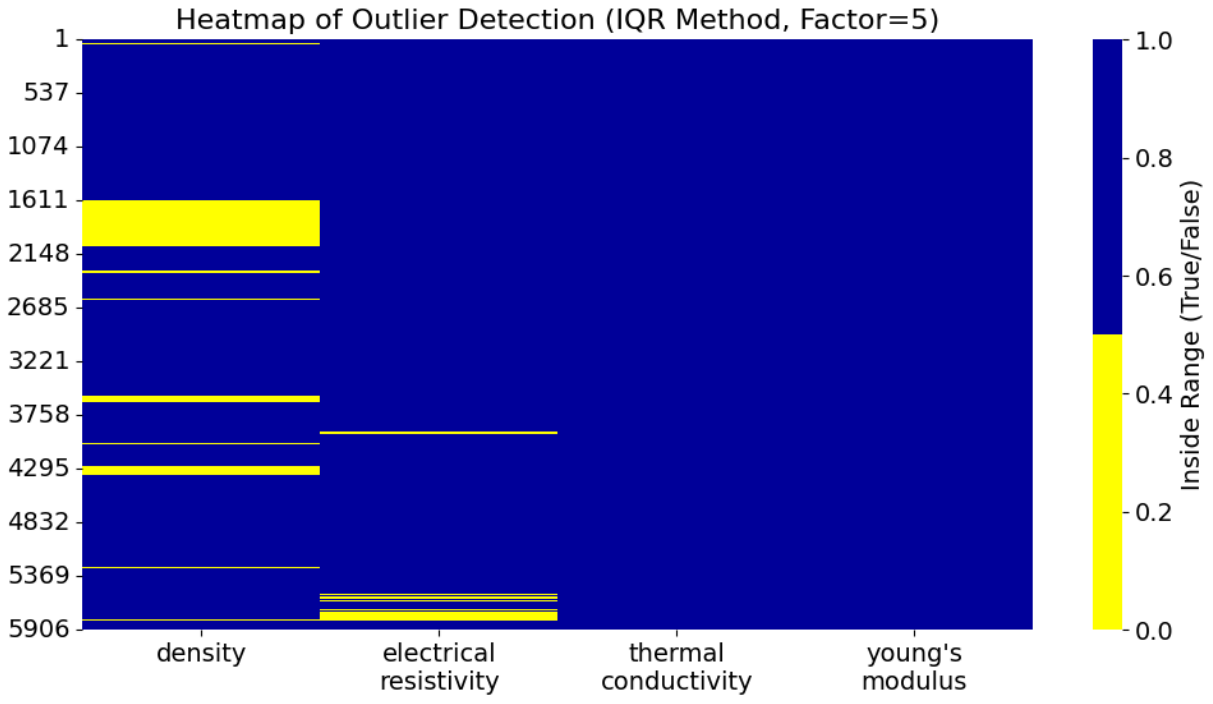

Рисунок 2 – Тепловая карта выбросов методом IQR для входных признаков
Fugure 2 – IQR emission heat map for input features

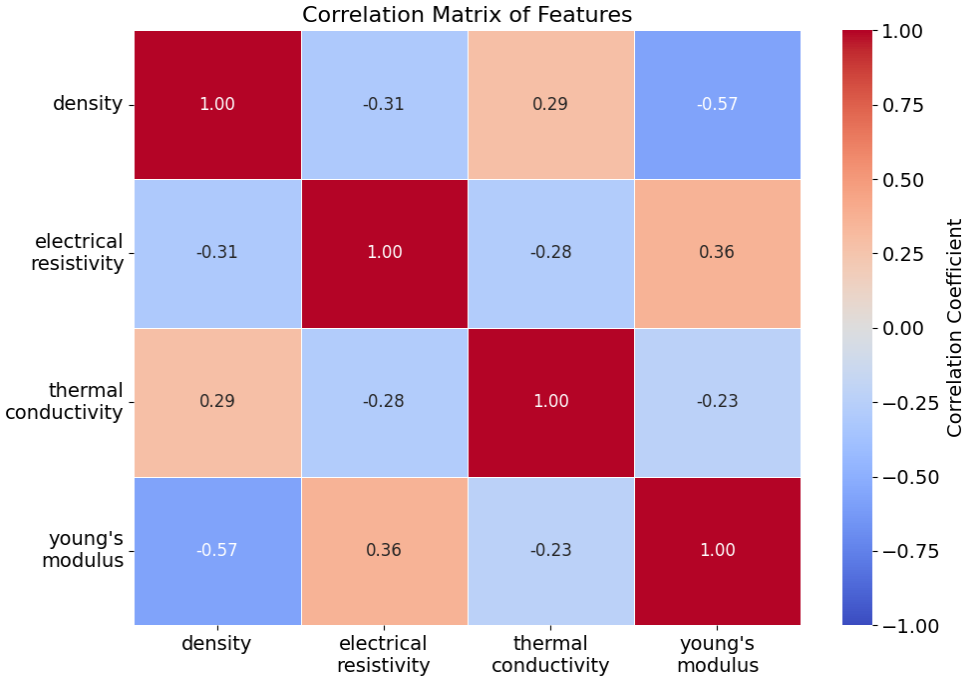

Рисунок 3 – Тепловая карта значения корреляции для входных признаков
Fugure 3 – Correlation value heat map for input features

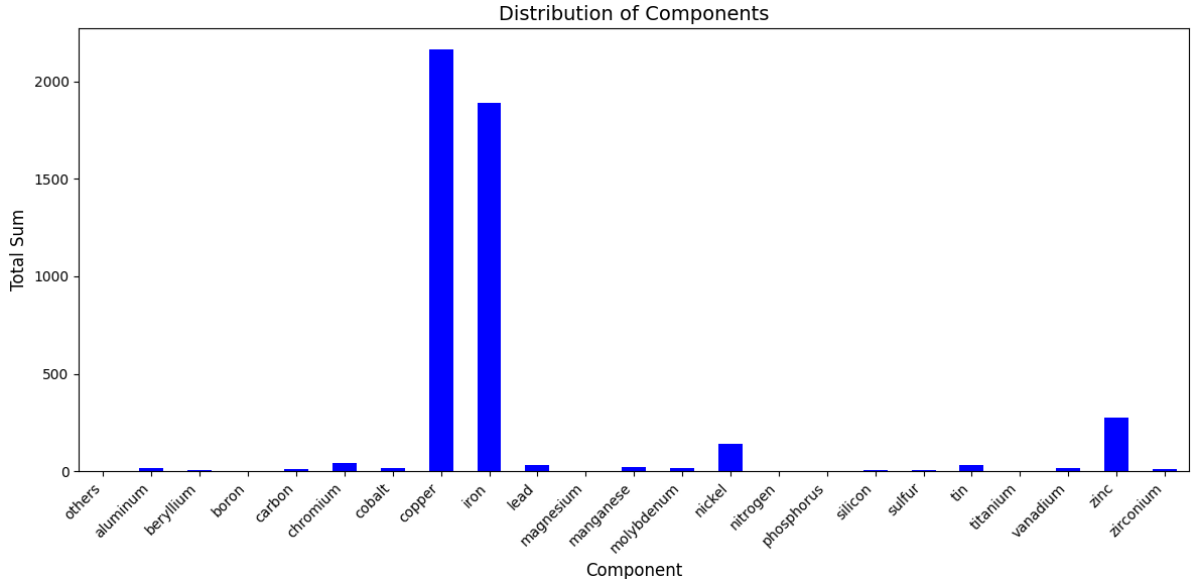

Рисунок 4 – Столбчатая диаграмма суммарных значений химических элементов
Figure 4 – Bar diagram of the total values of chemical elements

Полученные результаты на рисунке 4 демонстрируют, как различные химические элементы представлены в коллекции. Это послужит основой для более детального изучения взаимосвязей между их физическими свойствами и химическим составом.

На следующем этапе исходная коллекция данных была разделена на обучающую и тестовую выборки в соотношении 80/20. Такое разбиение обеспечивает возможность объективной оценки качества модели. Для оценки точности предсказаний используются метрики среднеквадратичной ошибки (MSE), средней абсолютной ошибки (MAE) и коэффициента детерминации (R^2). Эти метрики позволяют объективно сравнить различные модели и выбрать наиболее эффективную.

Предварительная обработка данных, включая очистку, нормализацию и разбиение на выборки, обеспечивает надёжную основу для построения и обучения моделей. Эти этапы гарантируют корректность и воспроизводимость результатов.

Для настройки начальных гиперпараметров нейронной сети были выбраны следующие значения:

– Количество нейронов в скрытых слоях: 64;

– Коэффициент обучения: 0.01;

– Количество слоёв: 2;

– Коэффициент исключения (dropout): 0.5.

Выбор этих параметров обоснован теоретическими предположениями, с учетом специфики задачи и размера набора данных. Для каждого из исследуемых методов и метрик построены столбчатые диаграммы на рисунке 5, что позволяет наглядно сравнить результаты, оценить их точность и выбрать оптимальные параметры.

Из анализа столбчатых диаграмм видно, что алгоритм случайного леса демонстрирует наименьшие значения средней абсолютной ошибки (MAE) и среднеквадратичной ошибки (MSE), что указывает на его способность эффективно минимизировать ошибку и на высокую точность результатов. Кроме того, случайный лес имеет наивысшее значение коэффициента детерминации (R^2), что свидетельствует о его высокой способности объяснять вариацию в данных.

Метод RidgeRegression также показывает неплохие результаты, но его производительность уступает случайному лесу. Нейронная сеть, наоборот, имеет отрицательное значение коэффициента детерминации (R^2), что указывает на то, что модель плохо адаптирована к данным и работает хуже, чем простая модель, предсказывающая постоянное значение. Это может быть связано с переобучением, недостаточной обучающей выборкой или неподходящими гиперпараметрами.

Для улучшения производительности нейронной сети был проведен поиск по гиперпараметрам, в результате которого были выбраны более оптимальные параметры:

– Количество нейронов в скрытых слоях: 128.

– Коэффициент обучения: 0.01.

– Количество скрытых слоев: 3.

– Коэффициент исключения (dropout): 0.0.

После настройки гиперпараметров для каждого из исследуемых методов и метрик были повторно построены столбчатые диаграммы, что позволило наглядно оценить изменения в производительности модели.

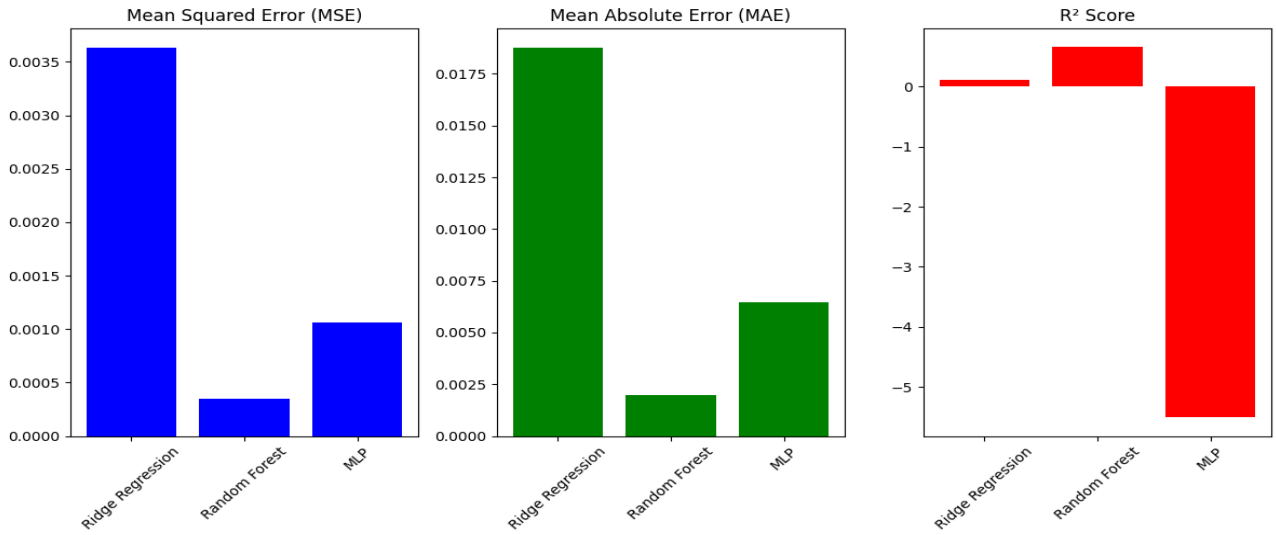

Рисунок 5 – Столбчатые диаграммы для среднеквадратичной ошибки, средней абсолютной ошибки и коэффициента детерминации для линейной регрессии, алгоритма случайного леса и нейронной сети с теоретическим подбором гиперпараметров

Figure 5 – Bar diagrams for mean square error, mean absolute error, and determination coefficient of linear regression, random forest algorithm, and neural network with theoretical hyperparameter selection

На диаграммах рисунка 6 видно, что производительность модели нейронной сети значительно улучшилась после оптимизации гиперпараметров. Увеличение числа нейронов и слоев позволило модели лучше захватывать и анализировать сложные нелинейные зависимости в данных. Снижение коэффициента исключения (dropout) способствовало сохранению большего числа активных нейронов во время обучения, что улучшило обучение модели, однако такой подход может привести к риску переобучения. Тем не менее ошибки модели уменьшились, что свидетельствует о повышении точности предсказаний.

Коэффициент детерминации (R^2) значительно увеличился и почти сравнялся с результатами метода случайного леса. Это означает, что модель стала гораздо лучше объяснять вариацию в данных. Однако, несмотря на эти улучшения, нейронная сеть все еще показывает немного худшие результаты, чем метод случайного леса. Это может быть связано с несколькими факторами:

1. **Размер и качество данных**. Нейронные сети требуют больших объемов данных для достижения высокой производительности. Если данные ограничены или содержат шум, многослойный перцептрон (MLP) может переобучиться или недообучиться, в то время как RandomForest лучше работает с ограниченными наборами данных и более устойчив к шуму.

2. **Гиперпараметры нейронной сети**. В текущем исследовании были рассмотрены только некоторые гиперпараметры. Для улучшения результатов стоит провести дополнительные эксперименты с другими гиперпараметрами, такими как функция потерь, функция активации и другие параметры. Эти аспекты будут исследованы в будущих работах.

3. **Кросс-валидация**. Необходимо провести анализ устойчивости методов к изменениям выборок данных для более точной оценки обобщающей способности моделей. Возможно, модель случайного леса не будет показывать таких высоких результатов на других выборках данных.

4. **Дисбаланс классов.** Если задача классификации или регрессии затронута дисбалансом классов или значениями выходных параметров, это может повлиять на способность модели эффективно обучаться. Например, в случае классификации редких классов или слабо представленных химических элементов модели могут не обучиться на этих примерах должным образом.

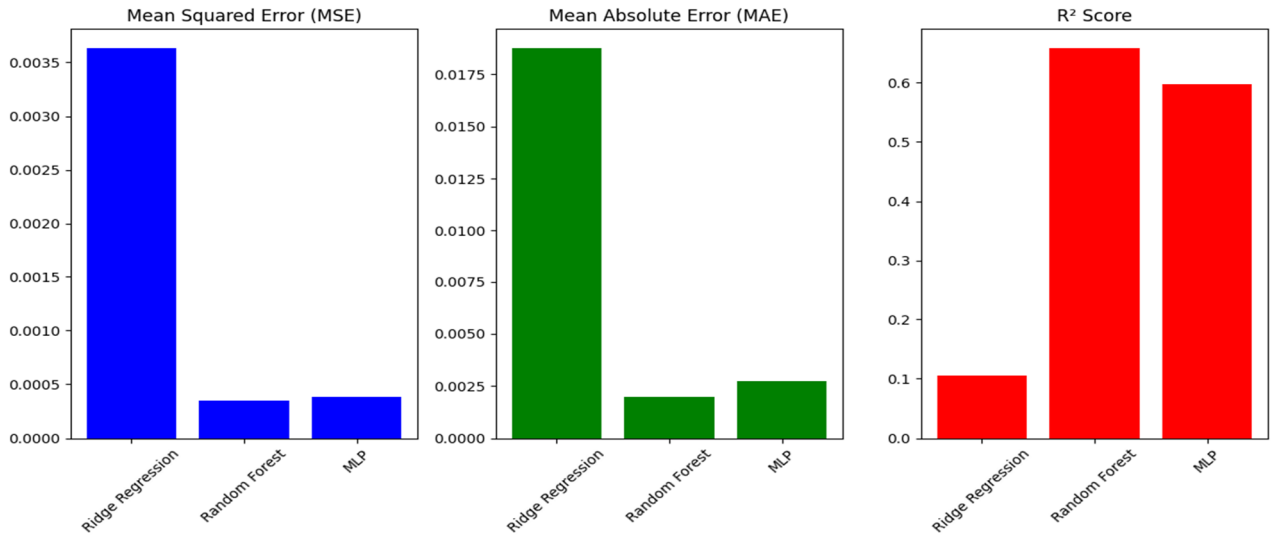

Рисунок 6 – Столбчатые диаграммы для среднеквадратичной ошибки, средней абсолютной ошибки и коэффициента детерминации для линейной регрессии, алгоритма случайного леса и нейронной сети после поиска оптимальных гиперпараметров
Figure 6 – Bar diagrams for mean square error, mean absolute error, and coefficient of determination for linear regression, random forest algorithm, and neural network after optimal hyperparameter search

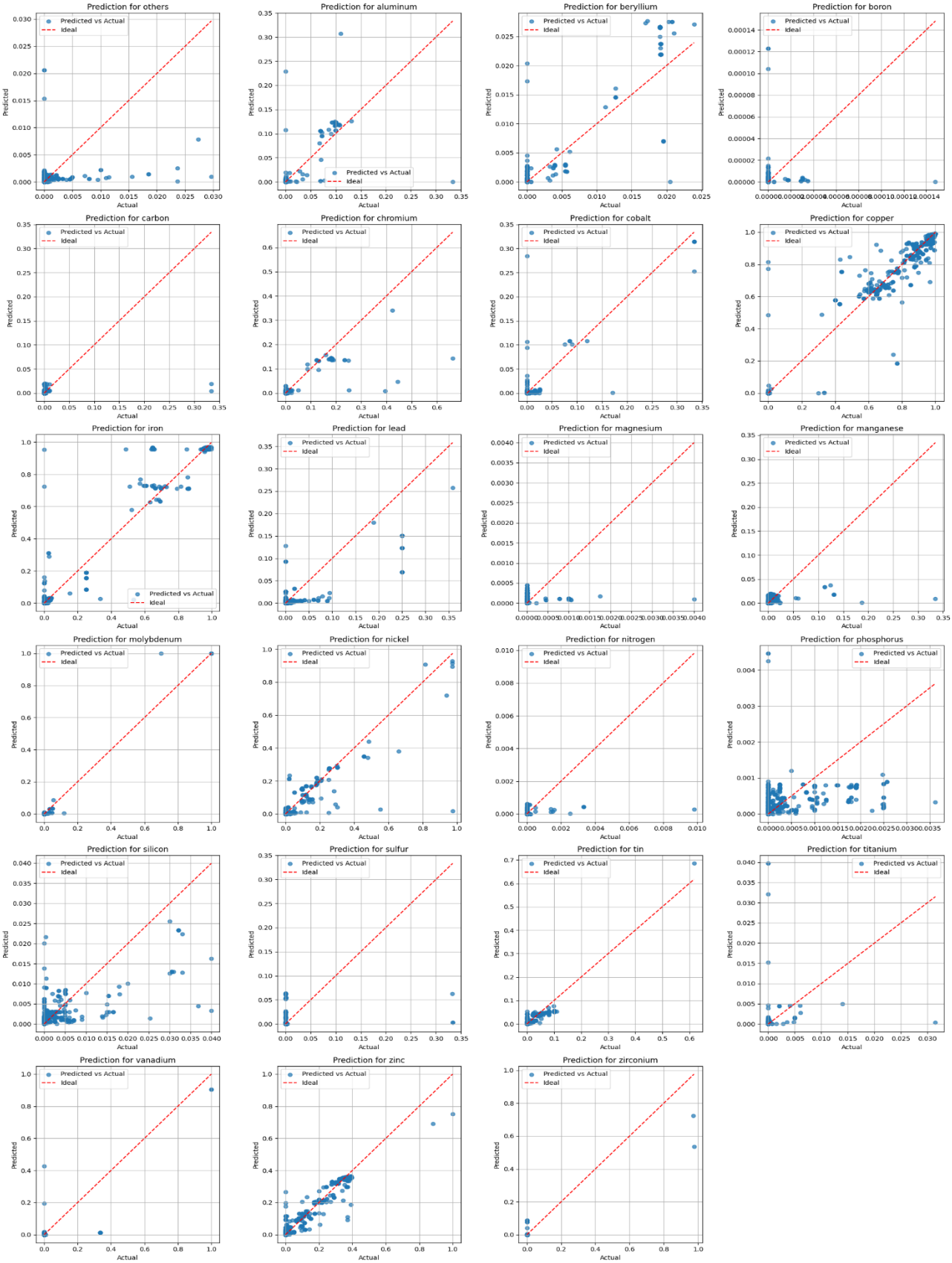

Рисунок 7– Диаграммы рассеивания для каждого химического элемента выходного слоя нейронной сети после подбора гиперпараметров
Figure 7 – Scattering diagrams for each chemical element of neural network output layer after hyperparame-ter selection

Далее для оценки качества модели будут визуализированы диаграммы рассеяния, показывающие зависимость между истинными значениями целевых параметров и предсказаниями модели для каждого из них. Такие графики ошибок позволяют наглядно оценить, насколько точно модель предсказывает значения различных параметров.

Из рисунка 7 можно заметить, насколько хорошо предсказания модели совпадают с истинными значениями. Идеальные предсказания должны располагаться вдоль красной прямой, которая отображает равенство истинных и предсказанных значений. Для наиболее распространенных компонентов материалов (медь, никель и другие) предсказания результаты имеют высокое качество. Однако, с уменьшением популярности химического компонента, уменьшается и его точность предсказаний. Что также говорит, о нарушении баланса данных.

Для оценки влияния признаков на предсказание модели была построена столбчатая диаграмма, которая позволяет выявить ключевые факторы для моделирования зависимости между входными данными и целевыми значениями.

Из рисунка 8 можно сделать выводы, что наиболее значимыми признаками являются модуль упругости и плотность. Это может указывать на наличие несбалансированности данных. В качестве оптимизации в будущих исследованиях стоит рассмотреть подходы к работе с несбалансированными наборами данных.

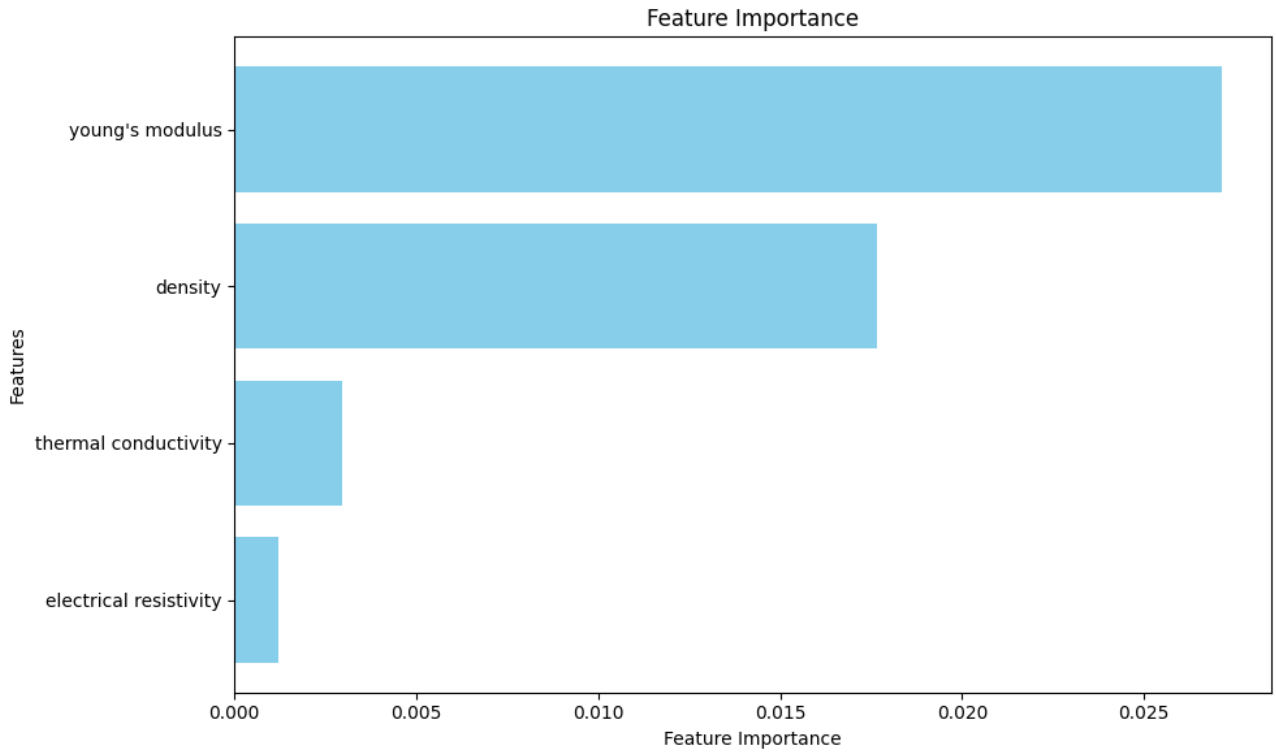

Рисунок 8 – Столбчатая диаграмма влияния признаков на выходные значения
Figure 8 – Bar diagram of features effect on input values

Заключение

В ходе проведенного исследования была разработана методика предсказания химического состава материалов на основе их физических свойств с использованием подхода многоклассовой регрессии. Это исследование стало важным шагом в направлении создания интеллектуальной системы, способной поддерживать процессы проектирования новых материалов, что может значительно ускорить разработку высококачественных материалов с оптимальными характеристиками.

Методика продемонстрировала свою эффективность в предсказании химического состава материалов, однако результаты могут быть улучшены. Для повышения точности предсказаний рекомендуется рассмотреть методы обработки несбалансированных данных, такие как

перераспределение классов или использование более сложных методов, например фокальной кросс-энтропии, которая помогает уменьшить влияние легко классифицируемых примеров. В рамках работы с нейронной сетью была проведена оптимизация гиперпараметров, однако протестированы лишь некоторые из них. В будущих исследованиях следует расширить перечень гиперпараметров, включая различные функции активации, методы регуляризации, а также внедрить кросс-валидацию для более точной оценки обобщающей способности моделей. Кроме того, полезным будет исследовать, какие дополнительные признаки могут улучшить точность модели, например классификация материалов или данные о процессах их обработки. Это позволит более глубоко понять факторы, влияющие на характеристики материалов.

Результаты данного исследования имеют значительный потенциал для разработки новых материалов, позволяя предсказать их химический состав на основе известных физических свойств. Это особенно важно в области материаловедения, где требуется быстрое и точное определение состава материалов для улучшения их характеристик. Применение данной методики поможет значительно сократить количество дорогостоящих экспериментальных исследований, что приведет к уменьшению затрат на разработку новых материалов. Потенциально эта методика может найти широкое применение в таких отраслях, как машиностроение, энергетика и другие области промышленности, где важен выбор оптимальных материалов для создания различных конструкций и деталей. Модели предсказания могут помочь в расчете стоимости материалов с оптимальными свойствами, что будет важным аспектом при оценке жизнеспособности разработки новых материалов для различных отраслей.

В дальнейшем планируется продолжить исследования в направлении повышения точности предсказаний, а также внедрить разработанную методику в практическую плоскость для разработки и поиска новых материалов с заданными характеристиками. Дополнительно методика будет адаптирована для работы с другими типами материалов, включая композиционные материалы и жидкости, что расширит возможности ее применения в различных областях науки и техники. В будущем также рассматривается возможность интеграции разработанной методики с другими задачами, такими как предсказание классификации [17], для создания единой системы моделирования взаимосвязей между различными характеристиками материалов в области информатики материалов.

Библиографический список

1. **Hautier G., Jain A., Ong S. P.** From the computer to the laboratory: materials discovery and design using first-principles calculations // Journal of Materials Science. 2012. Vol. 47. Pp. 7317-7340.

2. **Isayev O., Tropsha A., Curtarolo S. (ed.).** Materials informatics: methods, tools, and applications. John Wiley & Sons. 2019.

3. **Reiser P. et al.** Graph neural networks for materials science and chemistry // Communications Materials. 2022. Vol. 3. No. 1. P. 93.

4. **Isayev O., Tropsha A., Curtarolo S. (ed.).** Materials informatics: methods, tools, and applications // John Wiley & Sons, 2019.

5. **Zhang P., Li S. X., Zhang Z. F.** General relationship between strength and hardness // Materials Science and Engineering: A. 2011. Vol. 529. Pp. 62-73.

6. **Balachandran P.V., Theiler J., Rondinelli J.M., Lookman T.** Materials prediction via classification learning. Sci Rep. 5, 13285 (2015).

7. **Honysz R.** Modeling the chemical composition of ferritic stainless steels with the use of artificial neural networks // Metals. 2021. Vol. 11. No. 5. P. 724.

8. **Gaultois M.W. et al.** Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties // Apl Materials. 2016. Vol. 4. No. 5.

9. **Bayaga A.** Multinomial Logistic Regression: Usage and Application in Risk Analysis // Journal of applied quantitative methods. 2010. Vol. 5. No. 2.

10. **Liu Y., Wang Y., Zhang J.** New machine learning algorithm: Random forest // Information Computing and Applications: Third International Conference, ICICA 2012. Chengde, China, September 14-16, 2012. Proceedings 3. Springer Berlin Heidelberg. 2012. Pp. 246-252.

11. **Jais I.K.M., Ismail A.R., Nisa S.Q.** Adam optimization algorithm for wide and deep neural network // Knowl. Eng. Data Sci. 2019. Vol. 2. No. 1. Pp. 41-46.

12. **Gordon-Rodriguez E. et al.** Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. 2020.

13. **Liashchynskyi P., Liashchynskyi P.** Grid search, random search, genetic algorithm: a big comparison for NAS //arXiv preprint arXiv:1912.06059. 2019.

14. **Rahm E. et al.** Data cleaning: Problems and current approaches // IEEE Data Eng. Bull. 2000. Vol. 23. No. 4. Pp. 3-13.

15. **Dash M., Liu H., and Yao J.** Dimensionality reduction of unsupervised data // In Proc. 1997 IEEE Int. Conf. Tools with AI (ICTAI'97), pages 532-539, IEEE Computer Society, 1997.

16. **Dastjerdy B., Saeidi A., Heidarzadeh S.** Review of applicable outlier detection methods to treat geomechanical data // Geotechnics. 2023. Vol. 3. No. 2. Pp. 375-396.

17. **Корячко В.П., Викулин С.Д., Волков А.В.** Применение методов кластеризации для анализа свойств материалов // Вестник Рязанского государственного радиотехнического университета. 2024. № 89. С. 77-83.

UDC 004.724

COMPARATIVE STUDY OF MACHINE LEARNING METHODS AND NEURAL NETWORKS FOR PREDICTING CHEMICAL COMPOSITION OF MATERIALS

V. P. Koryachko, Dr. Sc. (Tech.), full professor, Head of the Department, RSREU, Ryazan, Russia;
orcid.org/0000-0000-0000-000X, e-mail: koryachko.v.p@rsreu.ru
S. D. Vikulin, post-graduate student, RSREU, Ryazan, Russia;
orcid.org/0009-0002-9932-1113, e-mail: vikulin97@gmail.ru
A. V. Volkov, specialist, BMSTU, Moscow, Russia;
orcid.org/0009-0008-1162-3816, e-mail: vic-volk@yandex.ru

The problem of developing methods for predicting the chemical composition of materials based on their physical properties using machine learning approaches is considered. ***The aim of this work*** *is to study the relationships between physical and chemical properties of materials to develop an intelligent system in order to support new materials design. Machine learning algorithms such as linear regression, decision trees, and neural networks to predict the chemical composition of different materials were employed. The study highlights the effectiveness of the proposed methods for accurately predicting the chemical composition, which can optimize material development process and improve material properties.*

Keywords: *machine learning, multi-class regression, chemical composition, physical properties of materials, neural networks, linear regression, decision trees, material science.*

DOI: 10.21667/1995-4565-2025-91-50-63

References

1. **Hautier G., Jain A., Ong S. P.** From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science*. 2012, vol. 47, pp. 7317-7340.

2. **Isayev O., Tropsha A., Curtarolo S. (ed.).** *Materials informatics: methods, tools, and applications*. John Wiley & Sons. 2019.

3. **Reiser P. et al.** Graph neural networks for materials science and chemistry. *Communications Materials*. 2022, vol. 3, no. 1, p. 93.

4. **Isayev O., Tropsha A., Curtarolo S. (ed.).** *Materials informatics: methods, tools, and applications*. John Wiley & Sons. 2019.

5. **Zhang P., Li S. X., Zhang Z. F.** General relationship between strength and hardness. *Materials Science and Engineering: A*. 2011, vol. 529, pp. 62-73.

6. **Balachandran P.V., Theiler J., Rondinelli J.M., Lookman T.** Materials prediction via classification learning. *Sci Rep*. 5, 13285 (2015).

7. **Honysz R.** Modeling the chemical composition of ferritic stainless steels with the use of artificial neural networks. *Metals*. 2021, vol. 11, no. 5, p. 724.

8. **Gaultois M.W. et al.** Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *Apl Materials*. 2016, vol. 4, no. 5.

9. **Bayaga A.** Multinomial Logistic Regression: Usage and Application in Risk Analysis. *Journal of applied quantitative methods*. 2010, vol. 5, no. 2.

10. **Liu Y., Wang Y., Zhang J.** New machine learning algorithm: Random forest. *Information Computing and Applications: Third International Conference, ICICA 2012*. Chengde, China, September 14-16, 2012. Proceedings 3. Springer Berlin Heidelberg. 2012, pp. 246-252.

11. **Jais I.K.M., Ismail A.R., Nisa S.Q.** Adam optimization algorithm for wide and deep neural network. *Knowl. Eng. Data Sci*. 2019, vol. 2, no. 1. pp. 41-46.

12. **Gordon-Rodriguez E. et al.** *Uses and abuses of the cross-entropy loss: Case studies in modern deep learning*. 2020.

13. **Liashchynskyi P., Liashchynskyi P.** Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv*:1912.06059. 2019.

14. **Rahm E. et al.** Data cleaning: Problems and current approaches. *IEEE Data Eng*. Bull. 2000, vol. 23, no. 4, pp. 3-13.

15. **Dash M., Liu H., and Yao J.** Dimensionality reduction of unsupervised data. In Proc. 1997 *IEEE Int. Conf. Tools with AI (ICTAI'97),* pages 532-539, IEEE Computer Society, 1997.

16. **Dastjerdy B., Saeidi A., Heidarzadeh S.** Review of applicable outlier detection methods to treat geomechanical data. *Geotechnics*. 2023, vol. 3, no. 2, pp. 375-396.

17. **Koryachko V.P., Vikulin S.D., Volkov A.V.** Primenenie metodov klasterizacii dlya analiza svojstv materialov. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2024, no. 89, pp. 77-83. (in Russian).