

ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ
И ТЕЛЕКОММУНИКАЦИИ

INFORMATION
TECHNOLOGIES
AND TELECOMMUNICATIONS

УДК 004.023+004.891.3

DOI: 10.17212/2782-2001-2025-2-93-110

Метод анализа данных на основе итерационной предварительной обработки и ансамблевых моделей для классификации технического состояния силовых трансформаторов*

А.И. ХАЛЬЯСМАА

620062, РФ, г. Екатеринбург, ул. Мира, 19, Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина

a.i.khaliasmaa@urfu.ru

При решении задач трансформации и анализа информации важным этапом является предварительная обработка данных. В системах обработки и анализа данных, поступающих от различных источников, особенно при невысоком уровне автоматизации процессов сбора данных и неполном охвате оборудования системами мониторинга, наборы данных, как правило, содержат большое количество пропущенных значений, а также могут содержать выбросы и ошибки формата. Исключение записей или признаков с большим относительным количеством пропущенных значений может привести к слишком высоким потерям данных. В задачах оценки технического состояния оборудования дополнительной особенностью, не позволяющей исключать большое количество записей, является несбалансированность классов, так как записей, относящихся к неудовлетворительному техническому состоянию, на 1–2 порядка меньше, чем относящихся к удовлетворительному и хорошему состоянию. Поэтому актуальной является разработка методов, позволяющих при предварительной обработке сохранять как можно больший объем исходных данных. В качестве примера рассмотрена оценка технического состояния силовых маслонаполненных трансформаторов. Наличие множества методов диагностики, которые основаны на учете процессов различной физической природы, требует агрегированного анализа данных. В качестве алгоритма предварительной обработки данных предложен итерационный алгоритм выбора признаков и заполнения пропущенных значений. Алгоритм позволил повысить степень заполненности с 75 до 94 %. Таким образом, количество пропусков, которые потребовалось заполнить синтетическими значениями, сократилось в 2,5 раза. Обработанная выборка данных была использована для построения моделей машинного обучения, выполняющих

* Статья получена 20 января 2025 г.

Работа выполнена в рамках государственного задания при финансовой поддержке Министерства науки и высшего образования Российской Федерации (тема № FEUZ-2025-0005. Разработка моделей и методов объяснимого искусственного интеллекта для повышения надежности и безопасности внедрения распределенных интеллектуальных систем на объектах электроэнергетики).

классификацию технического состояния трансформаторов. Для учета многоклассовости и несбалансированности выборки использовано построение двух последовательных классификаторов. Итоговая точность классификации по метрике $F1$ составила 83 %.

Ключевые слова: предварительная обработка данных, ансамблевые модели, машинное обучение, классификация, заполнение пропущенных значений, силовой трансформатор, техническое состояние оборудования, диагностика оборудования

ВВЕДЕНИЕ

При анализе наборов данных, полученных из множества различных источников, требуется особое внимание уделять предобработке данных, поскольку качество и полнота входных данных оказывают прямое влияние на точность последующего анализа. Отсутствие систематизированного подхода к обработке данных может приводить к неверным результатам оценки, а следовательно, к повышению рисков аварийных ситуаций и недоотпуску электрической энергии.

Одна из наиболее распространенных проблем в анализе данных о техническом состоянии электрооборудования – наличие пропущенных значений. Причинами таких пропусков могут быть как технические сбои измерительных устройств, так и ошибки в передаче данных; кроме того, для формирования датасетов с достаточной глубиной ретроспективы используются данные из рукописных журналов, которые изначально могут содержать ошибочные значения и еще большее количество искажений, появившихся в процессе оцифровки.

Примером оборудования, для оценки технического состояния (ТС) которого используются различные методы диагностики, является высоковольтный силовой маслонаполненный трансформатор. Такие трансформаторы выступают не только важными узлами энергетических сетей, но и объектами с высокой стоимостью и длительным сроком службы. Их эксплуатация сопряжена с воздействием значительных электрических нагрузок и факторов внешней среды [1]. Поэтому актуальной становится задача своевременного и точного определения ТС данного типа оборудования. Традиционные подходы, основанные на планово-предупредительном обслуживании и ремонте, постепенно уступают место подходу обслуживания по фактическому техническому состоянию с применением систем мониторинга и анализа данных. Такой подход позволяет выявлять признаки дефектов на ранней стадии и минимизировать как риски выхода оборудования из строя, так и затраты на ремонт.

Для оценки ТС маслонаполненных силовых трансформаторов используются различные методы, такие как анализ трансформаторного масла [2–4], обнаружение частичных разрядов [5, 6], тепловизионный контроль [7], оценка сопротивления обмоток и коэффициента трансформации и другие [8–10]. Широкое применение цифровых систем сбора и передачи данных в электроэнергетике позволило реализовать системы оперативной диагностики высоковольтного оборудования [11, 12].

В то же время возрастающий объем данных и необходимость оперативного принятия решений приводят к возможности решения задач оценки и классификации ТС силовых трансформаторов в автоматизированном режиме на основе анализа данных, полученных различными методами диагностики. Сложность и многообразие процессов различной физической природы, протекающих

в силовом маслонаполненном трансформаторе, делают классификацию его ТС сложной задачей, связанной с аппроксимацией нелинейной многофакторной зависимости [13]. Широко используются модели на базе машинного обучения, анализирующие результаты хроматографического анализа растворенных в трансформаторном масле газов [14, 15]. Применяются также нечеткий логический вывод [16] и ансамбли деревьев решений [17, 18], при этом используются как результаты хроматографического анализа, так и прочие признаки.

Современные подходы к обработке пропущенных данных включают использование статистических методов, машинного обучения и интеллектуального анализа данных. В статье [19] представлен обзор методов обработки пропущенных данных, включая их классификацию, преимущества и ограничения. Значительное внимание в научных исследованиях уделяется проблеме восстановления пропущенных значений в данных временных рядов [20, 21], поскольку такие данные широко используются в инженерных, экономических и физических задачах. В простейшем случае восстановление может быть выполнено с использованием методов интерполяции при наличии лишь одного признака. Однако более сложные случаи, связанные с многомерными временными рядами и зависимыми данными, требуют применения прогнозных моделей, таких как рекуррентные нейронные сети, авторегрессионные модели и методы машинного обучения.

Выбор конкретного метода восполнения пропущенных данных определяется природой данных, их структурными особенностями и характерными закономерностями, свойственными рассматриваемой технической системе. При этом важно учитывать не только точность восстановления, но и уровень неопределенности восстановленных значений, поскольку ошибки на этом этапе могут значительно повлиять на результаты дальнейших расчетов и моделирования. В связи с этим во многих прикладных задачах предпочтение отдается гибридным подходам, комбинирующим статистические методы с алгоритмами машинного обучения, что позволяет достичь оптимального баланса между точностью и обобщающей способностью модели.

1. ПОСТАНОВКА ЗАДАЧИ

Задача обработки пропущенных значений, как правило, состоит в минимизации их количества при сохранении как можно большего объема исходных данных и в общем виде может быть записана следующим образом:

$$D_b \leftarrow T(D_a), \quad (1)$$

$$M(D_b) < M(D_a), \quad (2)$$

$$M(D_b) \rightarrow \min, \quad (3)$$

$$V(D_b) \rightarrow \max, \quad (4)$$

где D_a – набор данных до предобработки; D_b – набор данных после предобработки; T – оператор предобработки; M – функция подсчета количества пропущенных значений; V – функция подсчета количества заполненных значений.

Кроме выражений (1) – (4) постановка задачи должна включать критерий, характеризующий сохранение статистических свойств набора данных, иначе

задача могла бы тривиально решаться заполнением пропусков нулями. Таким критерием можно быть количество пропущенных значений, заполненных синтетическими (сгенерированными) значениями.

Из постановки задачи видна противоречивость критериев, поскольку удаление записей или признаков с пропусками приведет к минимальному количеству заполненных пропусков, но и общий объем используемых исходных, несинтезированных данных может существенно снизиться. Синтезированные данные для заполнения пропусков отчасти преобразуют датасет в смоделированный. С учетом отмеченной выше несбалансированности классов высок риск, что модель, обученная на синтезированных данных в записях малочисленных классов, будет некорректно обрабатывать реальные данные соответствующих классов.

Конечные критерии качества выполнения предварительной обработки должны соответствовать решаемой задаче. Например, для задачи классификации следует использовать такие показатели, как точность, полнота, гармоническое среднее и другие, для задачи регрессии – метрики, основанные на вычислении отклонений выходных значений модели от фактических данных.

Цель исследования – разработка алгоритма предварительной обработки данных с учетом большого количества пропущенных значений и необходимости минимизации объема синтетических данных, а также создание модели на базе машинного обучения для решения задачи классификации ТС силовых маслонаполненных трансформаторов с учетом многоклассовости и несбалансированности классов. В качестве прикладной задачи рассмотрена оценка ТС силового трансформатора. Она является задачей многоклассовой классификации на основе анализа табличных данных (диагностических признаков).

2. МЕТОДОЛОГИЯ

2.1. АЛГОРИТМ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Обобщенная схема разработанного алгоритма показана на рис. 1. Количество итераций предобработки определяется свойствами набора данных, в большинстве случаев требуется выполнение двух итераций. Алгоритм одной итерации показан на рис. 2. Далее в тексте используются следующие обозначения: D – набор данных в формате плоской таблицы (двумерной матрицы), $cut_r(D, r)$ – исключение из D строк (записей), номера которых содержит множество r ; $cut_c(D, c)$ – исключение из D столбцов (признаков), номера которых содержит множество c ; n_i , m_i – количество записей и признаков в D_k .

1. Исключение заведомо неинформативных признаков u_1 :

$$D_1 \leftarrow cut_c(D, u_1). \quad (5)$$

2. Выполнение синтаксического анализа данных для исправления форматных несоответствий, включая избыточные символы, вариативность символов десятичного разделителя и другие отклонения от ожидаемого формата:

$$D_2 \leftarrow p(D_1). \quad (6)$$

Для отсечения выбросов используется величина IQR :

$$IQR = Q_3 - Q_1, \quad (7)$$

С отсечением записей, имеющих значения вне интервала:

$$Q_1 - 1,5IQR : Q_3 + 1,5IQR.$$

Выбросы и значения, не поддающиеся исправлениям формата, заменяются на обозначение NaN (Not a Number), которое поддерживается как особый вариант значения численных данных в базах данных и языках программирования.



Рис. 1. Алгоритм классификации с итерационной предобработкой

Fig. 1. An algorithm for classifying with iterative preprocessing

3. Исключение признаков, содержащих значительный объем отсутствующих данных:

$$D_{g3} \leftarrow (D_2 = \text{NaN}). \quad (8)$$

Подсчет количества пропущенных значений по признакам и исключение столбцов:

$$CG_i \leftarrow \frac{1}{n_2} \sum_{j=1}^{n_2} D_{g3ji}, i=1, \dots, m_2, \quad (9)$$

$$u_3 \leftarrow (CG > Th_c), \quad (10)$$

$$D_3 \leftarrow cut_{col}(D_2, u_3), \quad (11)$$

где Th_c – порог отсечения.

4. Исключение записей делается аналогично шагу 3.

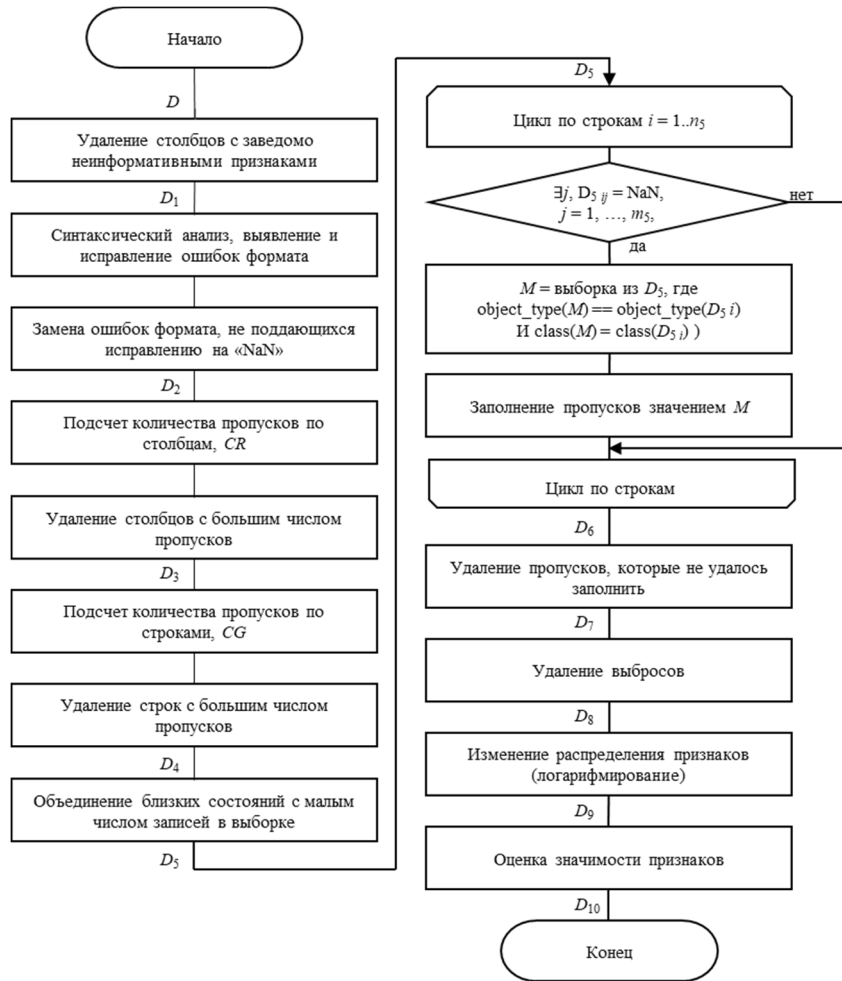


Рис. 2. Алгоритм одной итерации предобработки

Fig. 2. A flow chart of the data preprocessing iteration

5. Консолидация классов (при необходимости), характеризующих схожие состояния, для уменьшения разнородности данных и повышения устойчивости классификации:

$$D_5 \leftarrow j(D_4, Th_{join}). \quad (12)$$

6. Заполнение пропущенных значений. Выполняется итерационный процесс по записям выборки. Если в записи есть отсутствующее значение, оно заменяется медианным значением признака, рассчитанным по объектам того же типа и класса ТС. Если замена невозможна, пропуски остаются незаполненными.

7. Исключение записей с пропусками:

$$D_{g7} \leftarrow (D_6 = \text{NaN}). \quad (13)$$

$$RG_i \leftarrow \frac{1}{m} \sum_{j=1}^m D_{g7ij}, \quad i = 1, \dots, n_6. \quad (14)$$

$$u_7 \leftarrow (RG < 0), \quad (15)$$

$$D_7 \leftarrow \text{cut}_{\text{row}}(D_6, u_7). \quad (16)$$

8. Исключение выбросов:

$$u_8 \leftarrow (D_7 < b_{\text{lower}} \text{ ИЛИ } D_7 > b_{\text{upper}}), \quad (17)$$

$$D_8 \leftarrow \text{cut}_{\text{row}}(D_7, u_8), \quad (18)$$

где $b_{\text{lower}}, b_{\text{upper}}$ – векторы граничных значений по каждому признаку.

9. Логарифмирование признаков:

$$D_9 \leftarrow \log(D_{8ij} + \varepsilon) j \in T, \quad i = 1, \dots, n_8, \quad j = 1, \dots, m_8, \quad (19)$$

где ε – малое число (на 4 порядка меньше минимального значения признака).

10. Корреляционный анализ и построение классификационной модели на основе ансамбля решающих деревьев. Для выявления групп взаимосвязанных признаков, которые можно заменить одним из группы, используется коэффициент корреляции Спирмена.

Кроме того, выполняется обучение модели $\varphi(D_9)$, в результате чего формируется оценка значимости признаков, чтобы убедиться, что признаки действительно могут быть исключены. Этот шаг направлен на оптимизацию признакового пространства с сохранением высокой точности модели.

Формирование итогового набора данных после первой итерации:

$$D_{10} \leftarrow \text{cut}_{\text{col}}(D_9, u_{10}). \quad (20)$$

На этапе удаления пропущенных значений могли быть устранены записи, содержащие отсутствующие данные по признакам, которые затем определены как избыточные или слабоинформативные. Поэтому после их исключения необходимо пересмотреть пороги отсечения для восстановления ранее удаленных записей.

Применение такого подхода способствует увеличению объема выборки, используемой для обучения, поскольку предотвращает удаление записей, содержащих пропущенные значения признаков, которые ранее оказались исключены.

На второй итерации процесс отличается тем, что шаги 1–2 не выполняются, а в качестве исходного набора данных используется D_{10} первой итерации. Двух итераций, как правило, оказывается достаточно для достижения требуемого качества данных и точности модели, построенной на них.

2.2. ПОСТРОЕНИЕ МОДЕЛЕЙ ДЛЯ КЛАССИФИКАЦИИ ТС

Рассматриваемая задача характеризуется несбалансированностью в отношении количества записей, принадлежащих различным классам. Значимость ложноположительных и ложноотрицательных ошибок также различна. В связи с этим классификационный процесс организован в два этапа при использовании каскадного подхода.

На первом этапе формируется классификатор C_1 для разделения на неудовлетворительное и удовлетворительное/хорошее ТС. Его задача – надежное выявление записей с неудовлетворительным ТС, поскольку ошибки в этой категории являются наиболее значимыми. На втором этапе создается классификатор C_2 , как показано на рис. 3.

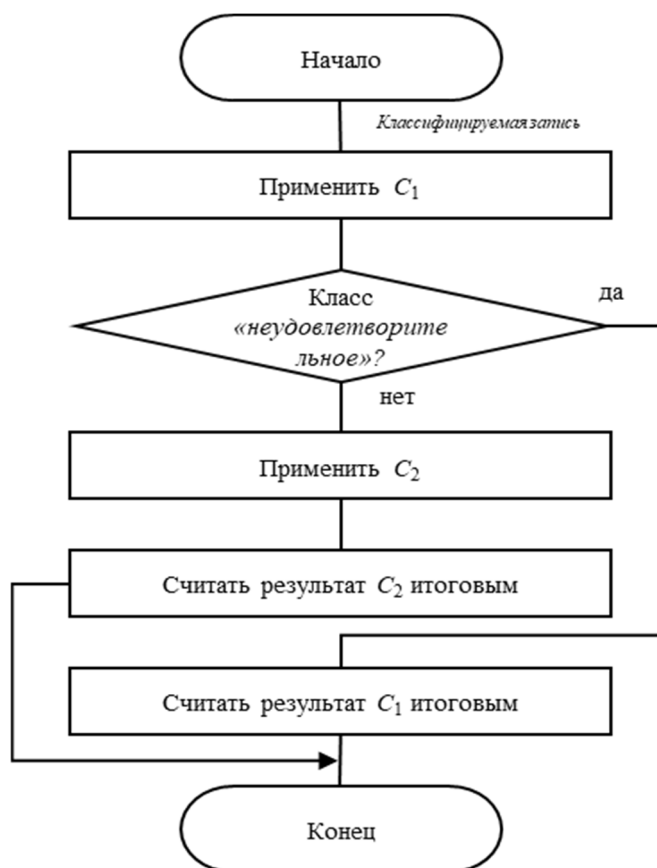


Рис. 3. Алгоритм многоклассовой классификации ТС

Fig. 3. A flow chart of multiclass classification of technical state

Такой последовательный подход с двумя классификаторами позволяет минимизировать вероятность критически важных ошибок и повысить точность классификации. Для оценки классификатора используются:

- точность (positive predictive values, PPV),
- полнота (true positive rate, TPR),
- гармоническое среднее между ними (F_1).

3. РЕЗУЛЬТАТЫ

3.1. ИСХОДНАЯ ВЫБОРКА И РЕЗУЛЬТАТЫ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ

Собранный автором массив данных представляет собой совокупность диагностических результатов и оценок технического состояния 365 силовых маслонаполненных трансформаторов класса напряжения 110 кВ. Данный набор включает 731 наблюдение (запись), каждое из которых описано 44 признаковыми характеристиками, приведенными в табл. 1 [18], где указана также доля пропущенных значений по каждому из признаков. Обозначения НН, СН, ВН в табл. 1 означают низкое, среднее и высокое напряжение.

Таблица 1

Table 1

Признаки исходной выборки данных

Features of the initial data sample

Признак		Доля пропущенных значений, %
Паспортные данные		
Диспетчерское наименование		0
Марка трансформатора		0
Вид переключающего устройства		0
Класс напряжения		0
Физико-химические характеристики масла		
Год производства масла		33
Пробивное напряжение, кВ		6
Влагосодержание, г/т		40
Кислотное число, мгКОН/г		25
Тангенс масла при 90 °С, %		11
Концентрация присадки ионол, %		72
Температура вспышки в закрытом тигле, °С		37
Хроматографический анализ масла в баке трансформатора		
Н ₂ , %		20
СН ₄ , %		9
С ₂ Н ₄ , %		9
С ₂ Н ₆ , %		18
С ₂ Н ₂ , %		9
СО ₂ , %		9
СО, %		11
Характеристики магнитной системы		
Потери холостого хода в пусконаладочных испытаниях при закороченной фазе, Вт	A	48
	B	48
	C	48
Потери холостого хода в последних испытаниях при закороченной фазе, Вт	A	48
	B	48
	C	48

Окончание табл. 1

End of the Tab. 1

Признак		Доля пропущенных значений, %
Характеристики обмоток и устройства устройство регулирования напряжения под нагрузкой (РПН)		
Год выпуска обмоток		33
Год выпуска РПН		33
Пробивное напряжение масла из бака РПН, кВ		33
Температура обмоток при пусконаладочных испытаниях, °С		46
Температура обмоток при последних испытаниях, °С		27
Сопротивление изоляции (R_{60}) обмотки при пусконаладке, МОм	ВН	34
	СН	40
	НН	34
Тангенс изоляции обмотки при пусконаладке, %	ВН	37
	СН	42
	НН	37
R_{60} обмотки при последних испытаниях, МОм	ВН	20
	СН	22
	НН	22
Тангенс изоляции обмотки при последних испытаниях, %	ВН	20
	СН	40
	НН	26
Год выпуска РПН		33

Периодичность проведения обследований рассматриваемого оборудования варьируется в диапазоне от полугода до двух лет, что обусловлено регламентами технического обслуживания и эксплуатационными условиями.

Были исключены следующие признаки как заведомо неинформативные:

- диспетчерское наименование;
- класс напряжения (поскольку для всех записей класс напряжения 110 кВ).

Затем были исключены признаки с высоким количеством пропущенных значений:

- концентрация присадки ионол;
- температура обмоток при пусконаладочных испытаниях;
- все признаки по магнитной системе.

Распределение ТС до и после удаления признаков и экземпляров с высоким уровнем пропусков представлено на рис. 4.

Классы ТС «непригодное» и «неудовлетворительное» были объединены в один класс ТС «непригодное» по причине близкого значения при малом количестве экземпляров. Далее корреляционный анализ в соответствии с шагом 10 позволил определить ряд признаков, которые могут быть исключены, поскольку имеют высокий коэффициент корреляции с другими:

- год производства масла и год выпуска РПН соответствуют году выпуска трансформатора;
- признаки, характеризующие обмотки, имеют высокие коэффициенты корреляции для вводов высокого, среднего и низкого напряжений (0,76...0,87), поэтому оставлены соответствующие признаки только для ВН.

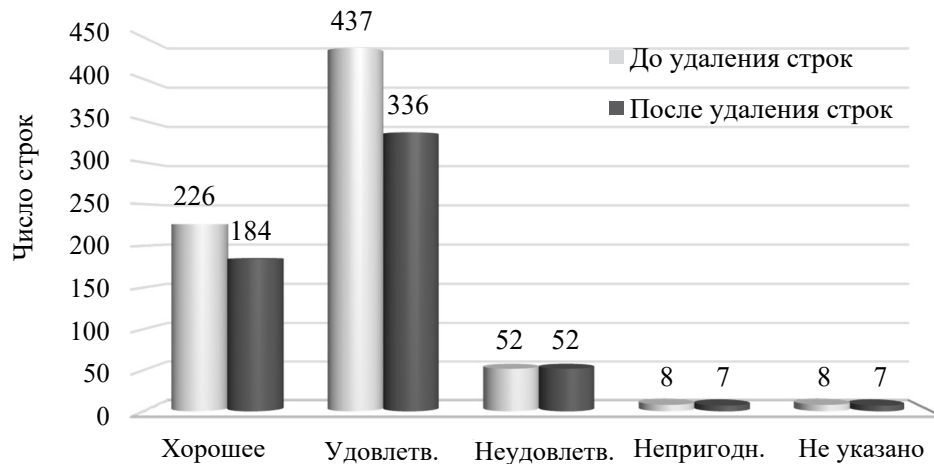


Рис. 4. Распределение данных по классам ТС

Fig. 4. Distribution of data according to technical state classes

Изначальная степень заполненности набора данных составляла 75 %, что свидетельствует о наличии значительного объема пропущенных значений в исходном массиве данных. Для повышения качества данных и уменьшения доли отсутствующих значений была проведена описанная выше процедура отбора информативных признаков. В результате двух итераций предобработки удалось повысить средний уровень заполненности выборки до 84 %.

Предложенный подход к предварительной обработке данных позволил минимизировать потери информации, особенно для объектов, относящихся к классу ТС «хорошее». В частности, благодаря применению двухитерационного метода очистки данных удалось сохранить в выборке на 9 % больше объектов данного класса.

Дополнительно за счет оптимизации стратегии работы с пропущенными значениями удалось увеличить общий объем данных на 5 % по всем классам, что способствует более надежному обучению моделей машинного обучения и улучшению качества оценки ТС. По завершении второй итерации предварительной обработки степень заполненности истинными значениями достигла 94 %.

При этом минимальный уровень заполненности по отдельным признакам увеличился до 87 %, что существенно выше исходных значений и обеспечивает более надежную основу для последующего анализа и построения моделей.

3.2. ПОСТРОЕНИЕ МОДЕЛИ КЛАССИФИКАЦИИ ТЕХНИЧЕСКОГО СОСТОЯНИЯ ТРАНСФОРМАТОРОВ

При формировании классификаторов C_1 и C_2 были использованы методы машинного обучения различной природы (метод k -ближайших соседей основан на вычислении расстояний в пространстве признаков, метод опорных векторов – на нелинейном преобразовании пространства признаков и построении оптимальной разделяющей гиперплоскости, ансамбли деревьев решений – на иерархических системах бинарных правил, сравнивающих признаки с пороговыми значениями). Результаты перечисленных методов приведены в табл. 2.

Реализации алгоритмов обучения моделей на языке программирования Python взяты из открытых библиотек [22–24].

Наилучшие показатели качества оказались у экстремального градиентного бустинга, в табл. 3 и 4 представлены его итоговые показатели на тестовом наборе данных. На рис. 5 показаны значимости признаков, а в табл. 5 итоговая матрица ошибок.

Таблица 2

Table 2

Результаты на тестовой выборке

Results for the test data sample

Метод	C ₁			C ₂		
	PPV, %	TPR, %	F ₁ , %	PPV, %	TPR, %	F ₁ , %
Метод опорных векторов	87	90	88	75	90	82
Метод <i>k</i> -ближайших соседей	100	90	95	77	89	82
Категориальный бустинг	100	90	95	70	100	83
Адаптивный бустинг	100	90	95	83	84	84
Случайный лес	100	80	89	78	96	86
Экстремальный градиентный бустинг	100	90	95	88	82	85

Таблица 3

Table 3

Результаты при использовании единого классификатора и двух последовательных

Results for using a single classifier and two consecutive ones

ТС	F ₁ , один классификатор, %	F ₁ , два классификатора, C ₁ и C ₂ %
Хорошее	65	75
Удовлетворительное	81	85
Неудовлетворительное	82	95
Средневзвешенное значение	76	81

Таблица 4

Table 4

Результаты использования разработанного метода классификации

Results of using the developed classification method

Показатель	Одна итерация предобработки	Две итерации предобработки
Количество признаков	18	24
Количество экземпляров	305	349
Средний процент заполненности, %	92	84
TPR, %	71	84
PPV, %	69	78

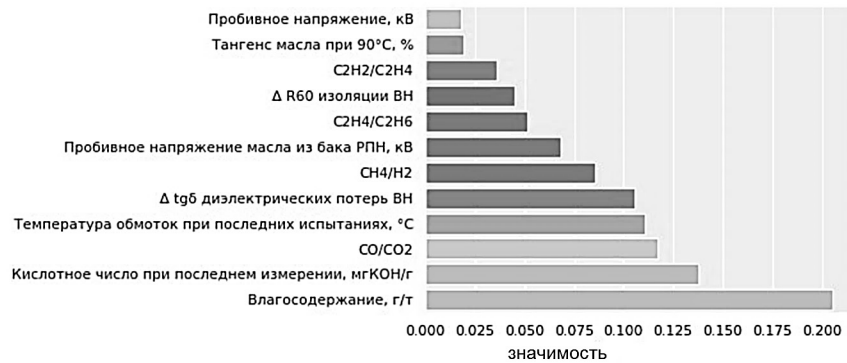


Рис. 7. Диаграмма значимости признаков

Fig. 7. Feature importance diagram

Таблица 5

Table 5

Результаты использования разработанного метода классификации**Results of using the developed classification method**

Фактическое ТС	ТС согласно результату классификации		
	Хорошее	Удовлетворительное	Неудовлетворительное
Хорошее	23 %	6 %	0 %
Удовлетворительное	10 %	47 %	0 %
Неудовлетворительное	0 %	1 %	13 %

Достоверность полученных значимостей признаков подтверждается эмпирическим опытом в области диагностики маслonaполненного трансформаторного оборудования.

Кроме того, было проведено статистическое сравнение результатов единого и двойного классификаторов на основе теста Макнемара:

$$\chi^2 = \frac{(|b - c| - 1)^2}{(b + c)}, \quad (21)$$

где b – количество экземпляров тестового набора данных, на которых единый классификатор допустил ошибку, а последовательность двух классификаторов – нет; c – количество обратных ситуаций.

Тест Макнемара – статистический метод, используемый для оценки разницы в долях ошибок между двумя классификационными моделями на одном и том же наборе данных. Он позволяет проверить, превосходит ли одна модель другую и является ли это статистически значимым.

По результатам теста $b = 17$, $c = 6$, $\chi^2_1 = 4,348$, p -значение = 0,03684. Обычно принимается уровень значимости $\alpha = 0,05$. Поскольку значение = 0,03684 < 0,05, статистически следует отвергнуть нулевую гипотезу, что означает существенные различия между единым классификатором и двумя последовательными. Эти результаты подтверждают, что различия между двумя классификаторами не являются случайными, а один классификатор показывает статистически значимое преимущество над другим.

ЗАКЛЮЧЕНИЕ

Рассмотрена задача обработки и анализа наборов данных с высоким количеством пропущенных значений. Разработанная итерационная процедура исключения малозначимых признаков позволяет на этапе предобработки выявить признаки, которые могут быть удалены. Такой подход способствует увеличению используемости имеющихся в таблице записей для целей классификации ТС, так как предотвращает удаление записей, содержащих пропуски по признакам, которые являются малоинформативными.

В ходе этапа предварительной обработки данных удалось существенно повысить степень заполненности набора данных, увеличив этот показатель с 75 до 94 %. Такой результат свидетельствует о значительном снижении объема пропущенных значений, что, в свою очередь, позволило уменьшить потребность в их замещении синтетическими данными. В частности, общее количество пропусков, требующих заполнения путем генерации синтетических данных, было сокращено в 2,5 раза, что позитивно сказалось на качестве выборки и потенциальной точности последующего анализа.

Такое увеличение заполненности данных стало возможным благодаря применению оптимизированных методов предобработки, включающих удаление наиболее проблемных записей, обработку пропущенных значений с учетом закономерностей в данных, что способствует улучшению качества моделей машинного обучения и повышает достоверность аналитических выводов.

Для классификации ТС силовых трансформаторов наилучшие результаты показала модель на основе экстремального градиентного бустинга. Разработанная модель состоит из двух классификаторов: первый выделяет наиболее важный класс неудовлетворительного ТС, второй разделяет оставшиеся трансформаторы на удовлетворительные и хорошие. В результате качество классификации по критерию гармонического среднего составило 81 % по сравнению с 76 % при использовании единого классификатора. Статистическая значимость различий между классификаторами была подтверждена с использованием теста Макнемара.

Экспериментально показано, что использование итерационной процедуры предварительной обработки данных повышает показатели классификации ТС. Точность классификации возросла с 69 до 78 %, а полнота увеличилась с 71 до 84 %, что свидетельствует о повышении качества модели и ее способности корректно идентифицировать классы объектов.

Разработанный метод может быть интегрирован в автоматизированные системы мониторинга технического состояния силовых трансформаторов, что позволит повысить надежность диагностики и снизить затраты на эксплуатацию оборудования. Будущие исследования могут быть направлены на адаптацию модели к данным, поступающим в режиме реального времени. Программная реализация моделей на языке Python с использованием только открытых библиотек позволяет реализовать программный модуль, который будет легко интегрироваться в существующие системы поддержки принятия решений.

Несмотря на достигнутые улучшения, метод может требовать дополнительной настройки при изменении характеристик входных данных. В частности, обработка дисбаланса классов остается важной задачей, требующей дальнейшего изучения. В перспективе возможно расширение метода на другие типы высоковольтного электрооборудования с учетом особенностей их диагностики.

СПИСОК ЛИТЕРАТУРЫ

1. *Davidenko I.V., Ovchinnikov K.V.* Identification of transformer defects via analyzing gases dissolved in oil // Russian Electrical Engineering. – 2019. – Vol. 90 (4). – P. 338–343. – DOI: 10.3103/S1068371219040035.
2. Hybrid DGA Method for power transformer faults diagnosis based on evolutionary k-means clustering and dissolved gas subsets analysis / A. Nanfak, S. Eke, F. Meghnefi, I. Fofana, G.M. Ngaleu, C.H. Kom // IEEE Transactions on Dielectrics and Electrical Insulation. – 2023. – Vol. 30 (5). – P. 2421–2428. – DOI: 10.1109/TDEI.2023.3275119.
3. Deep machine learning-based asset management approach for oil-immersed power transformers using dissolved gas analysis / L. Jin, D. Kim, K.Y. Chan, A. Abu-Siada // IEEE Access. – 2024. – Vol. 12. – P. 27794–27809. – DOI: 10.1109/ACCESS.2024.3366905.
4. *Misbahulmunir S., Ramachandaramurthy V.K., Thayoob Y.H.M.* Improved self-organizing map clustering of power transformer dissolved gas analysis using inputs pre-processing // IEEE Access. – 2020. – Vol. 8. – P. 71798–71811. – DOI: 10.1109/ACCESS.2020.2986726.
5. Identification and application of machine learning algorithms for transformer dissolved gas analysis / U.M. Rao, I. Fofana, K.N.V.P.S. Rajesh, and P. Picher // IEEE Transactions on Dielectrics and Electrical Insulation. – 2021. – Vol. 28 (5). – P. 1828–1835. – DOI: 10.1109/TDEI.2021.009770.
6. *Hussain M.R., Refaat S.S., Abu-Rub H.* Overview and partial discharge analysis of power transformers: a literature review // IEEE Access. – 2021. – Vol. 9. – P. 64587–64605. – DOI: 10.1109/ACCESS.2021.3075288.
7. *Fanchiang K.-H., Huang Y.-C., Kuo C.-C.* Power electric transformer fault diagnosis based on infrared thermal images using wasserstein generative adversarial networks and deep learning classifier // Electronics. – 2021. – Vol. 10 (10). – P. 1161. – DOI: 10.3390/electronics10101161.
8. Review of transformer health index from the perspective of survivability and condition assessment / S. Li, X. Li, Y. Cui, H. Li // Electronics. – 2023. – Vol. 12 (11). – P. 2407. – DOI: 10.3390/electronics12112407.
9. Deep learning in high voltage engineering: a literature review / S. Mantach, A. Lutfi, H. Moradi Tavasani, A. Ashraf, A. El-Hag, B. Kordi // Energies. – 2022. – Vol. 15 (14). – P. 5005. – DOI: 10.3390/en15145005.
10. *Хальясмаа А.И., Матренин П.В., Ерошенко С.А.* Оценка технического состояния силовых трансформаторов с использованием методов объяснимого искусственного интеллекта // Проблемы региональной энергетики. – 2024. – № 4 (64). – С. 1–9. – DOI: 10.52254/1857-0070.2024.4-64.01.
11. Contributions to monitoring the condition of substations / D. Sacerdotianu, F. Lazarescu, I. Hurezeanu, A.-M. Aciu, M. Nicola, I. Purcaru, A. Albata // Proceedings of the 2019 8th International Conference on Modern Power Systems (MPS), Cluj-Napoca, Cluj, Romania. – IEEE, 2019. – P. 1–6. – DOI: 10.1109/MPS.2019.8759689.
12. Power intelligent operation and maintenance system / Y. Ou, Z. Yan, T. Xie, N. Xi, H. Xie, X. Hu // Proceedings of the 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC), Rio De Janeiro, Brazil. – IEEE, 2023. – P. 23–27. – DOI: 10.1109/ICIRDC62824.2023.00010.
13. Power transformer fault diagnosis using neural network optimization techniques / V. Rokani, S.D. Kaminaris, P. Karaisas, D. Kaminaris // Mathematics. – 2023. – Vol. 11 (22). – P. 4693. – DOI: 10.3390/math11224693.
14. First measurement campaign by a multi-sensor robot for the lifecycle monitoring of transformers / J. Waikat, A. Jelidi, S. Lic, G. Sopidis, O. Kähler, A. Maly, J. Pestana, F. Fuhrmann, F. Belavić // Energies. – 2024. – Vol. 17 (5). – P. 1152. – DOI: 10.3390/en17051152.
15. Condition assessment of power transformers through DGA measurements evaluation using adaptive algorithms and deep learning / D.A. Barkas, S.D. Kaminaris, K.K. Kalkanis, G.Ch. Ioannidis, C.S. Psomopoulos // Energies. – 2023. – Vol. 16 (1). – P. 54. – DOI: 10.3390/en16010054.
16. *Ma H., Ekanayake C., Saha T.K.* Power transformer fault diagnosis under measurement originated uncertainties // IEEE Transactions on Dielectrics and Electrical Insulation. – 2012. – Vol. 19 (6). – P. 1982–1990. – DOI: 10.1109/TDEI.2012.6396956.

17. Khalyasmaa A., Senyuk M., Eroshenko S. Analysis of the state of high-voltage current transformers based on gradient boosting on decision trees // IEEE Transactions on Power Delivery and Electrical Insulation. – 2021. – Vol. 36 (4). – P. 2154–2163. – DOI: 10.1109/TPWRD.2020.3021702.
18. Data mining applied to decision support systems for power transformers' health diagnostics / A.I. Khalyasmaa, P.V. Matrenin, S.A. Eroshenko, V.Z. Manusov, A.M. Bramm, A.M. Romanov // Mathematics. – 2022. – Vol. 10 (14). – P. 2486. – DOI: 10.3390/math10142486.
19. Zhou Y., Arya S., Bouadjenek M.R. A comprehensive review of handling missing data: exploring special missing mechanisms // ArXiv. – 2024. – URL: <https://arxiv.org/html/2404.04905v1> (accessed: 21.05.2025).
20. Data imputation for multivariate time series sensor data with large gaps of missing data / R. Wu, S.D. Hamshaw, L. Yang, D.W. Kincaid, R. Etheridge, A. Ghasemkhani // IEEE Sensors Journal. – 2022. – Vol. 22 (11). – P. 10671–10683. – DOI: 10.1109/JSEN.2022.3166643.
21. Multi-type missing imputation of time-series power equipment monitoring data based on moving average filter–asymmetric denoising autoencoder / L. Jiang, J. Gu, X. Zhang, L. Hua, Y. Cai // Sensors. – 2023. – Vol. 23 (24). – P. 9697. – DOI: 10.3390/s23249697.
22. Scikit-Learn libraries. – URL: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed: 21.05.2025).
23. XGBoost. – URL: <https://github.com/dmlc/xgboost> (accessed: 21.05.2025).
24. CatBoost. – URL: <https://github.com/catboost> (accessed: 21.05.2025).

Халыясмаа Александра Ильмаровна, кандидат технических наук, доцент, заведующий научной лабораторией цифровых двойников в электроэнергетике, Уральский энергетический институт, Уральский федеральный университет имени первого Президента России Б.Н. Ельцина. Основное направление научных исследований – техническая диагностика электрооборудования, методы искусственного интеллекта в электроэнергетике, робототехника. Имеет более 250 печатных работ. E-mail: a.i.khaliasmaa@urfu.ru

Khalyasmaa Alexandra I., Ph.D (Eng.), Head of the scientific laboratory of digital twins in the power industry, Ural Power Engineering Institute, Ural Federal University, named after the first President of Russia, B.N. Yeltsin. The main field of her scientific research is technical diagnostics of electrical equipment and methods of artificial intelligence in power industry, robotics. She is the author of more than 250 scientific publications. E-mail: a.i.khaliasmaa@urfu.ru

DOI: 10.17212/2782-2001-2025-2-93-110

A data analysis method based on iterative preprocessing and ensemble models for classifying the technical condition of power transformers*

A.I. KHALYASMAA

Ural Federal University, named after the first President of Russia, B.N. Yeltsin, 19, Mira Street, Ekaterinburg, 620062, Russian Federation

a.i.khaliasmaa@urfu.ru

Abstract

Data preprocessing is a crucial stage in solving problems related to information transformation and analysis. In data processing and analysis systems that receive input from various sources (especially in cases where data collection processes have a low degree of automation and monitoring systems do not fully cover the equipment) datasets often contain a significant number of missing values, as well as potential outliers and format errors. Simply removing records or features with a high proportion of missing values can lead to an excessive data loss. In tasks

* Received 20 January 2025.

related to the assessment of the technical state of equipment, an additional factor preventing the removal of a large number of records is a class imbalance. Specifically, the number of records corresponding to an unsatisfactory technical condition is one to two orders of magnitude smaller than those representing satisfactory and good technical state. Therefore, it is crucial to develop preprocessing methods that retain as much of the original data as possible. As an example, we consider the assessment of the technical state of power oil-filled transformers. The existence of numerous diagnostic methods, each based on physical processes of different nature, necessitates an aggregated data analysis approach. To address this, an iterative feature selection and missing value imputation algorithm was proposed as a preprocessing method. This algorithm increased the data sets completeness from 75 % to 94 %, thereby reducing the number of missing values that required synthetic imputation by a factor of 2.5. The processed dataset was then used to train machine learning models designed to classify the transformer technical state. To account for both multi-class classification and dataset imbalance, a two-stage classification approach was implemented. The final F1-score of the classification model reached 83%, demonstrating the effectiveness of the preprocessing and modeling pipeline.

Keywords: data preprocessing, ensemble models, machine learning, classification, imputation of missing values, power transformer, equipment technical state, equipment diagnostics

REFERENCES

1. Davidenko I.V., Ovchinnikov K.V. Identification of transformer defects via analyzing gases dissolved in oil. *Russian Electrical Engineering*, 2019, vol. 90 (4), pp. 338–343. DOI: 10.3103/S1068371219040035.
2. Nanfak A., Eke S., Meghnefi F., Fofana I., Ngaleu G.M., Kom C.H. Hybrid DGA method for power transformer faults diagnosis based on evolutionary k-means clustering and dissolved gas subsets analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, 2023, vol. 30 (5), pp. 2421–2428. DOI: 10.1109/TDEI.2023.3275119.
3. Jin L., Kim D., Chan K.Y., Abu-Siada A. Deep machine learning-based asset management approach for oil-immersed power transformers using dissolved gas analysis. *IEEE Access*, 2024, vol. 12, pp. 27794–27809. DOI: 10.1109/ACCESS.2024.3366905.
4. Misbahulmunir S., Ramachandaramurthy V.K., Thayoob Y.H.M. Improved self-organizing map clustering of power transformer dissolved gas analysis using inputs pre-processing. *IEEE Access*, 2020, vol. 8, pp. 71798–71811. DOI: 10.1109/ACCESS.2020.2986726.
5. Rao U.M., Fofana I., Rajesh K.N.V.P.S., Picher P. Identification and application of machine learning algorithms for transformer dissolved gas analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, 2021, vol. 28 (5), pp. 1828–1835. DOI: 10.1109/TDEI.2021.009770.
6. Hussain M.R., Refaat S.S., Abu-Rub H. Overview and partial discharge analysis of power transformers: a literature review. *IEEE Access*, 2021, vol. 9, pp. 64587–64605. DOI: 10.1109/ACCESS.2021.3075288.
7. Fanchiang K.-H., Huang Y.-C., Kuo C.-C. Power electric transformer fault diagnosis based on infrared thermal images using wasserstein generative adversarial networks and deep learning classifier. *Electronics*, 2021, vol. 10 (10), p. 1161. DOI: 10.3390/electronics10101161.
8. Li S., Li X., Cui Y., Li H. Review of transformer health index from the perspective of survivability and condition assessment. *Electronics*, 2023, vol. 12 (11), p. 2407. DOI: 10.3390/electronics12112407.
9. Mantach S., Lutfi A., Moradi Tavasani H., Ashraf A., El-Hag A., Kordi B. Deep learning in high voltage engineering: a literature review. *Energies*, 2022, vol. 15 (14), p. 5005. DOI: 10.3390/en15145005.
10. Khalyasmaa A.I., Matrenin P.V., Eroshenko S.A. Assessment of power transformer technical state using explainable artificial intelligence. *Problemy regional'noi energetiki = Problems of the Regional Energetics*, 2024, no. 4 (64), pp. 1–9. DOI: 10.52254/1857-0070.2024.4-64.01. (In Russian).
11. Sacerdotianu D., Lazarescu F., Hurezeanu I., Aciu A.-M., Nicola M., Purcaru I., Albita A. Contributions to monitoring the condition of substations. *Proceedings of the 2019 8th International Conference on Modern Power Systems (MPS)*, Cluj-Napoca, Cluj, Romania, 2019, pp. 1–6. DOI: 10.1109/MPS.2019.8759689.

12. Ou Y., Yan Z., Xie T., Xi N., Xie H., Hu X. Power intelligent operation and maintenance system. *Proceedings of the 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC)*, Rio De Janeiro, Brazil, 2023, pp. 23–27. DOI: 10.1109/ICIRDC62824.2023.00010.
13. Rokani V., Kaminaris S.D., Karaisas P., Kaminaris D. Power transformer fault diagnosis using neural network optimization techniques. *Mathematics*, 2023, vol. 11 (22), p. 4693. DOI: 10.3390/math11224693.
14. Waikat J., Jelidi A., Lic S., Sopidis G., Kähler O., Maly A., Pestana J., Fuhrmann F., Belavić F. First measurement campaign by a multi-sensor robot for the lifecycle monitoring of transformers. *Energies*, 2024, vol. 17 (5), p. 1152. DOI: 10.3390/en17051152.
15. Barkas D.A., Kaminaris S.D., Kalkanis K.K., Ioannidis G.C., Psomopoulos C.S. Condition assessment of power transformers through dga measurements evaluation using adaptive algorithms and deep learning. *Energies*, 2023, vol. 16 (1), p. 54. DOI: 10.3390/en16010054.
16. Ma H., Ekanayake C., Saha T.K. Power transformer fault diagnosis under measurement originated uncertainties. *IEEE Transactions on Dielectrics and Electrical Insulation*, 2012, vol. 19 (6), pp. 1982–1990. DOI: 10.1109/TDEI.2012.6396956.
17. Khalyasmaa A., Senyuk M., Eroshenko S. Analysis of the state of high-voltage current transformers based on gradient boosting on decision trees. *IEEE Transactions on Power Delivery and Electrical Insulation*, 2021, vol. 36 (4), pp. 2154–2163. DOI: 10.1109/TPWRD.2020.3021702.
18. Khalyasmaa A.I., Matrenin P.V., Eroshenko S.A., Manusov V.Z., Bramm A.M., Romanov A.M. Data mining applied to decision support systems for power transformers' health diagnostics. *Mathematics*, 2022, vol. 10 (14), p. 2486. DOI: 10.3390/math10142486.
19. Zhou Y., Arya S., Bouadjenek M.R. A comprehensive review of handling missing data: exploring special missing mechanisms. *ArXiv*, 2024. Available at: <https://arxiv.org/html/2404.04905v1> (accessed 21.05.2025).
20. Wu R., Hamshaw S.D., Yang L., Kincaid D.W., Etheridge R., Ghasemkhani A. data imputation for multivariate time series sensor data with large gaps of missing data. *IEEE Sensors Journal*, 2022, vol. 22 (11), pp. 10671–10683. DOI: 10.1109/JSEN.2022.3166643.
21. Jiang L., Gu J., Zhang X., Hua L., Cai Y. Multi-type missing imputation of time-series power equipment monitoring data based on moving average filter-asymmetric denoising autoencoder. *Sensors*, 2023, vol. 23 (24), p. 9697. DOI: 10.3390/s23249697.
22. *Scikit-Learn libraries*. Available at: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed 21.05.2025).
23. *XGBoost*. Available at: <https://github.com/dmlc/xgboost> (accessed 21.05.2025).
24. *CatBoost*. Available at: <https://github.com/catboost> (accessed 21.05.2025).

Для цитирования:

Халыясмаа А.И. Метод анализа данных на основе итерационной предварительной обработки и ансамблевых моделей для классификации технического состояния силовых трансформаторов // Системы анализа и обработки данных. – 2025. – № 2 (98). – С. 93–110. – DOI: 10.17212/2782-2001-2025-2-93-110.

For citation:

Khalyasmaa A.I. Metod analiza dannykh na osnove iteratsionnoi predvari-tel'noi obrabotki i ansamblevykh modelei dlya klassifi-katsii tekhnicheskogo sostoyaniya silovykh transformatorov [A data analysis method based on iterative preprocessing and ensemble models for classifying the technical condition of power transformers]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2025, no. 2 (98), pp. 93–110. DOI: 10.17212/2782-2001-2025-2-93-110.