DOI: 10.24892/RIJIE/20250302

Методы машинного обучения в задачах прогнозирования остаточного ресурса оборудования: гибридный подход*

Альчаков В.В., Писарев В.И.
Севастопольский государственный университет
г. Севастополь, Российская Федерация
alchakov@sevsu.ru, pisarev.valsidalv@gmail.com

Аннотация. **В статье рассматривается задача прогнозирования остаточного срока службы силовых трансформаторов на основе анализа концентраций растворенных газов в трансформаторном масле. Исследование базируется на временных рядах измерений концентраций H_2, CO, C_2H_4 и C_2H_2 в течение продолжительного периода. Данные были предварительно обработаны, агрегированы и проанализированы с целью выявления закономерностей и факторов, влияющих на износ оборудования. Для прогнозирования остаточного ресурса трансформаторов (RUL) использованы методы машинного обучения, включая модели градиентного бустинга (XGBoost, LightGBM, CatBoost), деревья решений (RandomForest, ExtraTrees) и многослойные персептроны (MLP). Рассмотрены различные стратегии агрегации данных, в том числе классическое усреднение и автоматическое извлечение признаков с использованием TSFresh. Дополнительно проведена классификация признаков для повышения точности регрессионного прогнозирования. Итогом исследования стал гибридный подход, сочетающий классификацию и регрессию, применяемый в задачах прогноза остаточного срока службы исследуемого оборудования.**

Ключевые слова: **машинное обучение, гибридная модель, остаточный ресурс оборудования, регрессия, прогностическая модель.**

Введение

Особый интерес представляют исследования, направленные на совершенствование прогнозирования остаточного ресурса (Remaining Useful Life, RUL) оборудования в условиях сложных и изменяющихся данных. Важным направлением является разработка гибридных моделей, сочетающих традиционные статистические подходы и нейросетевые методы. Например, комбинированные алгоритмы, включающие модели Facebook Prophet для выделения линейных трендов, LSTM (Long-Short Term Memory) для анализа сложных нелинейных зависимостей демонстрируют высокую точность и адаптивность в задачах прогнозирования [1].

Исследования в данной области продолжают активно развиваться, предлагая все более точные и адаптивные методы прогнозирования остаточного ресурса оборудования. Многообразие методов машинного обучения для прогнозирования RUL подробно рассмотрено в следующих работах [2-9]. В совокупности, представленные работы подчеркивают значимость интеграции гибридных моделей прогнозирования на базе технологий машинного обучения в промышленные системы и демонстрируют потенциал повышения надежности и эффективности эксплуатации трансформаторов и другого оборудования.

Они востребованы в атомной и традиционной энергетике, авиации, машиностроении, транспорте и даже в электроэнергетических системах, где позволяют прогнозировать срок службы трансформаторов, газоизолированных распределительных устройств и других критически важных компонентов.

Восстановительный ремонт, техническое обслуживание на основе надежности и интеллектуальные системы управления ресурсами активно используют алгоритмы прогнозирования RUL для оптимизации затрат и повышения эффективности. Например, в трансформаторах прогнозирование остаточного ресурса учитывает не только общее состояние оборудования, но и деградацию отдельных его компонентов, таких как изоляционные материалы и охлаждающие жидкости. В сфере авиации и железнодорожного транспорта модели машинного обучения помогают прогнозировать износ деталей, снижая риски аварий.

Силовые трансформаторы являются критически важными компонентами атомных электростанций (АЭС), обеспечивающими преобразование и распределение электроэнергии. В настоящее время значительное количество таких трансформаторов эксплуатируется с превышением назначенного срока службы, что требует разработки эффективных методов мониторинга и диагностики. В связи с этим особую актуальность приобретает задача прогнозирования остаточного ресурса силовых трансформаторов методами машинного обучения. Данный подход позволяет не только своевременно выявлять возможные неисправности, но и оптимизировать процессы технического обслуживания на основе текущего состояния оборудования [10].

Теоретические аспекты прогнозирования RUL

Прогнозирование RUL может осуществляться различными методами, которые делятся на четыре основные группы: статистическая оценка, прогноз параметров, регрессионные модели и анализ схожести с историческими данными [11].

Метод статистической оценки базируется на анализе исторических данных о длительности работы оборудования до отказа, на основе накопленной статистики строится функция распределения времени до отказа, позволяющая оценить вероятностные характеристики остаточного ресурса оборудования.

* Статья публикуется по рекомендации программного комитета Всероссийской научно-технической конференции "Пром-Инжиниринг", https://icie-rus.org

Подход прогнозирования параметров основан на расчете значений технологических параметров (например, концентрации газов) до пересечения контрольных пределов, а основными типами моделей являются линейная и экспоненциальная деградации.

Задачи прогнозирования остаточного ресурса можно решать при помощи моделей регрессии, выделяя информативные признаки из временных рядов технологических параметров и строя по ним регрессионную модель.

Подход на основе схожести с историческими данными основан на сравнении текущего состояния оборудования с накопленными данными о прошлых отказах оборудования.

Силовые трансформаторы АЭС работают в условиях длительной эксплуатации (порядка 25 лет), что обуславливает жесткий контроль технического состояния с целью предотвращения аварийных ситуаций. Основным методом диагностики силовых трансформаторов является хроматографический анализ растворенных газов (CADG). Так, в процессе длительной работы оборудования в трансформаторном масле накапливаются газы различных типов, концентрация которых отражает развитие внутренних дефектов:

1) водород (H_2) – указывает на наличие электрических разрядов низкой энергии;
2) оксид углерода (CO) – свидетельствует о термическом разложении бумажной изоляции;
3) этилен (C_2H_4) – образуется при перегреве масла;
4) ацетилен (C_2H_2) – характерен для электрических разрядов высокой энергии.

Анализ динамики этих концентраций позволяет определять степень деградации трансформатора и прогнозировать его остаточный ресурс с целью проведения своевременного обслуживания.

Описание и предобработка набора данных

Источником данных служит открытый датасет “Power Transformers FDD and RUL” [12], содержащий измерения концентраций растворенных газов в трансформаторном масле, с лицензией типа CC0 открытого распространения – данные открыты общественности и могут быть свободно использованы без ограничений. В рамках данного исследования используется набор данных с *Kaggle* – онлайн-платформа для обмена и публикации наборов данных, содержащие наборы данных, по стандартам безопасности (KDD CUP 99) [13].

Исходный набор данных состоит из набора файлов формата *.csv*, каждый из которых содержит информацию о концентрациях четырех газов (H_2, CO, C_2H_4 и C_2H_2) в трансформаторном масле для каждого из 2100 трансформаторов и файла разметки с данными о результирующей переменной. Данные представлены в виде временных рядов, состоящих из 420 точек измерений и самих четырех измерений, где каждая точка измерений соответствует концентрации газа в определенный момент времени, а период между измерениями составляет 12 часов. Таким образом, каждый файл содержит данные за 210 дней наблюдений.

Разметка для каждого трансформатора выражена в виде числового значения, представляющего собой остаточный ресурс трансформатора на конец периода наблюдений, выраженный в точках измерений (1 единица представляет собой диапазон в 12 часов между измерениями).

Для эффективной и удобной работы с данными, файлы *.csv* были объединены в словарь *X_data* (пример итогового представления на базе первых двух трансформаторов отображен в табл. 1), где значениями выступили объекты датафрейма, содержащие данные о концентрациях газов для каждого трансформатора, а ключами – идентификаторы трансформаторов (с целью соответствия формату файла разметки (табл. 2)).

Таблица 1

Представление данных для первых двух трансформаторов в словаре *X_data*

Идентификатор	Данные				
	номер измерения	H_2	CO	C_2H4	C_2H_2
2_trans_497.csv	0	0,001202	0,029565	0,001069	0,000251
	...	...	...	...	...
	419	0,002294	0,042099	0,004630	0,000345
2_trans_483.csv	0	0,001875	0,030855	0,002613	0,000068
	...	...	...	...	...
	419	0,002764	0,037128	0,006520	0,000146

Таблица 2

Представление данных для первых двух трансформаторов в файле разметки

Идентификатор трансформатора	Остаточный ресурс трансформатора
2_trans_497.csv	550
2_trans_483.csv	1093

С точки зрения структуры, исследуемый набор данных относится к типу многомерных временных рядов с зависимой переменной в разметке, что делает его подходящим для применения различных методов машинного обучения, в том числе и посредством объединения подходов (создания гибридных моделей), комбинирующих различные алгоритмы и подходы к анализу данных.

В ходе дальнейшей предобработки и исследования набора данных, был проведен анализ временных рядов концентраций газов для выявления возможной сезонности двумя способами: визуальный анализ графиков временных рядов и определение сезонности при помощи автокорреляционной функции (ACF) временных рядов каждого из газов. Полученные графики указывают на отсутствие сезонности в исследуемом периоде в 210 дней, было отмечено лишь повышение концентрации газов с течением времени, что будет учитываться уже прогнозными моделями. При визуальном изучении графиков ACF, выбросов значений, которые могли бы указывать на сезонность, обнаружено не было. Графики имеют плавный нисходящий тренд, что указывает на постепенное снижение корреляции между значениями временного ряда в виду увеличения дистанции между этими значениями – влияние предыдущих значений временного ряда уменьшается с течением времени. Однако, следует учитывать, что данный период относительно

короткий по сравнению с общим сроком службы трансформаторов, который может составлять десятки лет, поэтому отсутствие явной сезонности в данных может быть обусловлено ограниченной длиной выборки и не исключает возможности наличия сезонных закономерностей в более длительных временных интервалах.

С целью анализа набора данных с целевыми метками было проведено исследование распределения прогнозных значений остаточного времени работы трансформаторов, выраженных в единицах измерения времени (1 единица = 12 часов). После разбиения значений на кластеры было определено неравномерное распределение значений RUL – достаточно большая часть значений (725 из 2100) сосредоточена в диапазоне между 1019–1093 единицами времени, что составляет 34,52% от их общего числа (рис. 1). Остальные значения RUL распределены по другим кластерам с меньшей плотностью. Подобное неравномерное распределение значений может оказать существенное влияние на эффективность работы регрессионных моделей, поскольку доминирование определенного кластера при обучении может привести к смещению модели в сторону соответствующих значений, что негативно сказывается на точности последующего прогнозирования значений оставшихся данных из набора. Для решения этой проблемы был предложен подход, основанный на сочетании классификации и регрессии.

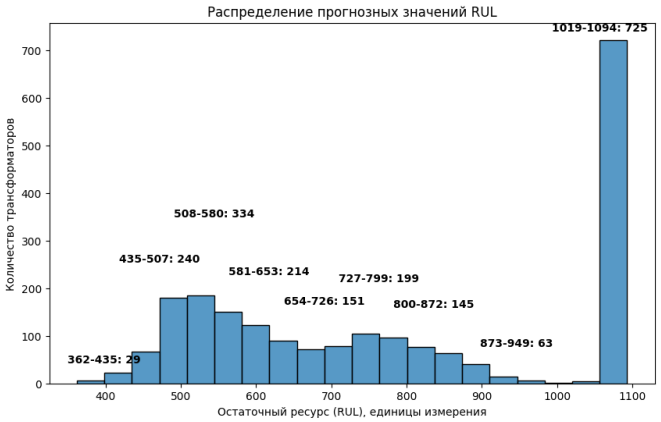

Рис. 1. Распределение прогнозных значений RUL

Такой подход позволит учесть специфику каждого кластера и повысить точность прогнозирования RUL для трансформаторов.

По итогам предварительной обработки данных, для проведения исследования, посвященного прогнозированию остаточного ресурса трансформаторов АЭС, был выбран подход, основанный на применении моделей регрессии, что обусловлено следующей спецификой датасета:

1) временные ряды отражают динамику концентрации газов в трансформаторном масле, позволяя использовать методы анализа временных рядов и извлекать признаки, релевантные для прогнозирования RUL;

2) отсутствуют данные о полных исторических пробегах трансформаторов до отказа, что ограничивает применение подходов, основанных на статистической оценке RUL или на схожести с паттернами из прошлого;

3) отсутствие четко определенных пороговых значений работоспособности оборудования затрудняет использование подходов, основанных на прогнозе пересечения контрольных пределов.

Далее перейдем к описанию методологии исследования.

Методология проведения исследования

Для решения задачи прогнозирования RUL трансформаторного оборудования был выбран подход, основанный на применении моделей регрессии на базе машинного обучения. Методология этого подхода включает методы агрегации данных, а также методы регрессии и классификации, которые также могут использоваться для формирования гибридных подходов, применяемые для повышения качества прогноза.

Методы агрегации данных.

Исходные данные, описанные выше, представляют собой словарь с набор временных рядов, отражающих динамику концентрации газов в трансформаторном масле для каждого отдельного трансформатора. Подобная структура датасета не позволяет напрямую применять методы машинного обучения для прогнозирования остаточного ресурса, поскольку используемые модели требуют представления данных в табличном виде с фиксированным количеством признаков для каждого объекта. В связи с этим, была выбрана стратегия агрегирования данных перед обучением моделей, преобразующая временные ряды в фиксированный набор статистических и временных характеристик. В качестве методов агрегации использовались классическая агрегация средним значением и автоматизированное извлечение признаков с помощью библиотеки *TSFresh*.

Так, простая агреграция средним была выбрана в качестве базового подхода, поскольку среднее значение является простым и интерпретируемым показателем, отражающим общую тенденцию изменения концентрации газов, что позволяет быстро привести данные к требуемому виду и оценить эффективность моделей машинного обучения на агрегированных признаках.

Библиотека *TSFresh* предназначена для автоматического извлечения признаков из временных рядов, предоставляя широкий спектр методов анализа временных рядов с целью выявления разнообразных характеристик исследуемых данных, а в ходе изучения эффективности разных подходов использовались две вариации набора параметров [14, 15]:

1) набор параметров *MinimalFCParametrs* предназначен для извлечения базовых признаков, таких как среднее значение, стандартное отклонение, квантили и пр., обеспечивая компактное представление временных рядов и позволяет оценить эффективность моделей машинного обучения на основе простых признаков (63 показателя);

2) набор параметров *EfficientFCParameters* применялся для извлечения более сложных признаков, которые оказались более информативными, обеспечивая возможность составления более сложных гибридных подходов. Параметры этого набора состоят из признаков, основанных на автокорреляции, спектральном анализе, энтропии и других методах анализа временных рядов (794 показателя).

Методы регрессии

Методы регрессии играют ключевую роль в задачах RUL, применяя их, строится математическая модель, связывающую наблюдаемые признаки (как в нашем случае, наборы параметров разной длины для концентрации газов в трансформаторном масле) с целевой переменной. Цель использования методов регрессии – получить модель, способную с достаточной точностью предсказывать целевую переменную на основе имеющихся данных, а последующий анализ позволяет выявить наиболее важные признаки, которые оказывают большее влияние на результирующую переменную.

В ходе исследования выбранные методы регрессии на были разделены на типы, что позволило систематизировать и сравнить различные способы построения регрессии к прогнозированию RUL. Кратко опишем их [16, 17].

Первым типом определены ансамблевые методы градиентного бустинга (*CatBoostRegressor*, *XGBRegressor* [18], Light*GBMRegressor* [19], *HistGradientBoostingRegressor*), основанные на последовательном построении деревьев решений, каждое из которых исправляет ошибки предыдущих, последовательно улучшая качество прогноза, несмотря на сложные нелинейные зависимости, а разница между текущим прогнозом модели и известным истинным значением целевой переменной называется остатком. Выбор этого способа построения регрессии был обусловлен многомерностью исследуемого набора данных, но в то же время, достаточно эффективной работы на ограниченных объемах данных, адаптируемостью регрессионной модели за счет возможности указания большого количества гиперпараметров, а также наличию встроенных механизмов работы с категориальными переменными, которые интегрируется в датасет в ходе процедуры классификации признаков. Однако, имеют достаточно высокую вычислительная сложность по сравнению с обычными деревьями решений, увеличивая время дообучения и построения прогноза, что в рамках нашей задачи не имеет существенного значения, поскольку данные с датчиков снимаются два раза в сутки.

Вторым типом определены ансамблевые методы построения деревьев решений, (*RandomForestRegressor*, *ExtraTreesRegressor*), основанный на построении ансамбля деревьев решений, суть которого заключается в обучении каждого дерева независимо от другого (в отличие от последовательного улучшения предсказаний, как в случае применения методов градиентного бустинга), но использует различные стратегии для повышения устойчивости модели к переобучению (что особенно важно в условиях исследований с ограниченным набором данных), выбросам и повышения точности прогнозирования [20].

Multi-Layer Perceptron Regressor – среди искусственных нейронных сетей прямого действия *MLPRegressor* является подклассом многослойных персептронов и используется для моделирования сложных нелинейных зависимостей между входными признаками и прогнозируемым значением. Однако, его высокая потребность в данных делает его менее устойчивым к ограниченным выборкам, характерным для данного исследования (1500 строк), повышая риск переобучения. В связи с этим, построение подобной модели требует тщательного подбора гиперпараметров, структуры сети, использование методов *l-1* и *l-2* регуляризации, а также дополнительного анализа для повышения интерпретируемости и минимизации риска переобучения, что в случае успеха позволит выявить скрытые нелинейные зависимости и повысить точность прогнозирования RUL.

Методы классификации признаков

Применение классификаторов в задачах прогнозирования RUL направлено на раннее выявление аномалий и повышение точности регрессионных моделей. Для построения гибридных моделей поочередно использовались два метода классификации признаков [20].

Логистическая регрессия выбрана за счет своей интерпретируемости, позволяя анализировать влияние отдельных признаков на вероятность скорого износа трансформатора (в случаях, когда прогнозируемое значение RUL меньше 1093 единиц времени). Несмотря на линейность, данный метод эффективен при умеренном количестве признаков, полученных в результате агрегации временных рядов.

RandomForestClassifier, будучи нелинейным ансамблевым методом, учитывает сложные зависимости между признаками, полученными в результате агрегации временных рядов (в особенности при агрегации на базе параметров *EfficientFCParameters*), и позволяет обнаруживать скрытые закономерности, недоступные линейным моделям, а устойчивость к переобучению и пропущенным значениям делает его надежным инструментом для применения в задачах прогнозирования RUL.

Так, рассмотренные методы классификации признаков позволяют исключить выбросы из обучающей выборки, что снижает искажения в моделях и улучшает качество предсказаний. Кроме того, классификаторы оптимизируют стратегию обслуживания, способствую своевременному выявлению оборудования, нуждающегося в обслуживании.

Статистические метрики оценки точности моделей

В рамках данного исследовании для оценки точности моделей регрессии и классификации применялось четыре статистические метрии, каждая из которых позволяет анализировать определенные критерии качества прогнозирования остаточного ресурса трансформаторного оборудования.

Так, $\bar{y}$ – среднее значение фактического, y_i – истинное значение (фактический RUL), $\hat{y}_i$ – предсказанное моделью значение, n – количество наблюдений.

Средняя абсолютная ошибка (*MAE*) определяет среднее абсолютное отклонение предсказанных значений от фактических, предоставляя интуитивно понятную интерпретацию ошибки – чем ближе к 0 значение средней абсолютной ошибки, тем точнее прогноз модели. Рассчитывается по формуле:

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

Коэффициент детерминации (R^2) оценивает, какая доля дисперсии зависимой переменной объясняется моделью. Значение метрики варьируется от 0 до 1 – чем ближе оно к 1, тем точнее модель описывает зависимости в данных. Рассчитывается по формуле:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (2)$$

Корень из среднеквадратичной ошибки (*RMSE*) измеряет среднее отклонение предсказанных значений от реальных, но при этом придает больший вес большим ошибкам, что достигается путем возведения разностей в квадрат, усреднения их и взятия квадратного корня. Такая методика делает корень из среднеквадратической ошибки более чувствительным к крупным отклонениям, в отличие от средней абсолютной ошибки, потому что их влияние на показатель сильнее. Рассчитывается по формуле:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Средняя абсолютная процентная ошибка (*MAPE*) определяет среднюю абсолютную ошибку прогнозирования в процентах от фактического значения. Метрика удобна для сравнения точности моделей на разных наборах данных, однако средняя ошибка предсказания от фактического RUL в процентном отношении также может характеризовать прогнозную модель. Рассчитывается по формуле:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

Метрика оценки качества классификации (*F1-score*) особенно полезна при несбалансированных классах, представляя собой гармоническое среднее между точностью (precision – доля объектов, предсказанных как принадлежащие к определенному классу, которые действительно принадлежат к нему) и полнотой (recall – доля объектов данного класса, которые модель корректно определила), объединяя их в единое значение. Высокое значение метрики указывает на баланс между точностью и полнотой и гарантирует, что модель не просто подстраивается под доминирующий класс, а качественно прогнозирует результирующую переменную. Рассчитывается по формуле:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

Описанные метрики будут использоваться для оценки качества моделей прогнозирования RUL [21].

Результаты исследования

В ходе работы рассматривались различные методы и составляющие прогнозирования оставшегося срока работоспособности трансформаторов – используются разные техники агрегации входных данных, классификации и последующего регрессионного анализа с целью определения прогнозного значения [22]. Так, в рамках исследования были определены 5 подходов к прогнозированию RUL, а главной задачей сравнения этих подходов являлось повышение точности прогнозирования за счет создания гибридных подходов из рассматриваемых методов и подбор гиперпараметров.

Подход №1: простая агрегация средним + модель регрессии

В рамках первого подхода метод агрегации данных заключался в вычислении средних значений признаков по каждому трансформатору, что позволило уменьшить размерность выборки и упростить ее структуру, однако подобное преобразование привело к потере временных зависимостей, негативно сказалось на точности предсказаний. Для построения модели регрессии были протестированы различные алгоритмы машинного обучения (описанные ранее), производится инференс модели.

Несмотря на явное переобучение моделей, наиболее эффективным методом оказался *XGBoostRegressor* (табл. 3), гиперпараметры модели не изменялись. Данный подход не является эффективным, поскольку агрегация данных простым средним значительно упрощает тренировочную выборку, что напрямую влияет на переобучение моделей.

Подход №2: TSFresh-агрегация (MinimalFCParametrs) + модель регрессии

С помощью библиотеки *TSFresh* (с набором параметров *MinimalFCParameters,* извлекающим базовые статистические характеристики временных рядов) производится извлечение признаков из временных рядов данных, описывающих концентрации газов для каждого трансформатора, данные стандартизируются (*StandartScaler*). Для построения модели регрессии были использованы алгоритмы машинного обучения из подхода № 1, производится инференс модели.

Использование автоматического извлечения признаков с помощью *TSFresh* позволило добиться существенного улучшения качества прогнозирования по сравнению с простой агрегацией данных. Применение модели регрессии на базе *LightGBM* на данном наборе признаков обеспечило значительное снижение средней абсолютной ошибки (в 1,83 раза) и увеличение коэффициента детерминации (в 2,14 раза), что подтверждает необходимость использования методов агрегации данных, учитывающих временные закономерности (табл. 3).

Однако, несмотря на улучшение результатов, методика *TSFresh* в базовой конфигурации ограничена набором извлекаемых статистических характеристик и не учитывает сложные нелинейные зависимости во временных рядах. Был сделан вывод о том, что дальнейшее повышение точности прогнозирования возможно при использовании более сложных подходов.

Подход №3: TSFresh-агрегация (MinimalFCParametrs) + классификация (логистическая регрессия) + модель регрессии

Аналогично подходу № 2, с помощью методов *TSFresh* (с набором параметров *MinimalFCParameters*) извлекаются признаки из временных рядов данных для каждого трансформатора, данные стандартизируются. На этапе классификации на основе извлеченных признаков применяется логистическая регрессия для разделения трансформаторов на две группы: "нормальные" и "аномальные" с точки зрения прогнозируемого RUL. В нашем случае аномальными считаются значения пикового остаточно ресурса – количество времени, равное 1093 единицам (RUL = 1093), затем алгоритм логистической регрессии определяет закономерности в данных, позволяющие отделить такие трансформаторы от остальных (присвоение значений, согласно классу – 0 и 1; этот параметр используется при построении модели регрессии). Качество этой классификации оценивается с помощью F1-меры. Происходит обучение моделей регрессии, инференс, оценка.

Наилучшие показатели демонстрирует модель на базе *LightGBM*, однако ее показатели оказались ниже, чем во втором подходе. Так, даже в условиях потери точности при построении регрессии, классификационный этап позволил отделить аномальные трансформаторы с качеством классификации 0,64 (табл. 3).

Несмотря на некоторое снижение численных показателей метрик гибридной модели по сравнению с аналогичным подходом без этапа предварительной классификации, наблюдается улучшение обобщающей способности модели, снижение переобучения и повышение качества модели, о чем также свидетельствует более точное соответствие распределения прогнозных значений фактическому распределению целевой переменной.

Таблица 3

Статистические метрики оценки качества моделей

Метрики	Подход № 1	Подход № 2	Подход № 3	Подход № 4
	XGBoost	*LightGBM*	*LightGBM*	*CatBoost*
F1_train	–	–	0,67	0,91
F1_test	–	–	0,64	0,79
MAE_train	136,77	31,29	71,27	38,16
MAE_test	160,9	87,65	103,56	84,67
R2_train	0,54	0,97	0,63	0,93
R2_test	0,35	0,75	0,54	0,71
RMSE_train	166,1	42,89	148,53	65,42
RMSE_test	194,98	121,05	164,45	129,54
MAPE_train, %	34,37	37,61	8,59	37,21
MAPE_test, %	34,27	36,43	13,24	36,12

Метрики	Подход № 5. Агрегация TSFresh (EfficientFCParameters) + классификация (RandomForestClassifier) + модель регрессии – лучшим в данном подходе определен ExtraTrees						
	Cat Boost	*XG Boost*	*Light GBM*	*HG Boost*	*Random Forest*	*Extra Trees*	*MLP*
F1_train	0,89	0,89	0,89	0,89	0,89	0,89	0,89
F1_test	0,82	0,82	0,82	0,82	0,82	0,82	0,82
MAE_train	45,6	29,36	25,97	32,5	31,09	33,7	28,41
MAE_test	71,5	68,89	66,87	67	68,89	66,77	97,19
R2_train	0,93	0,97	0,97	0,96	0,96	0,96	0,96
R2_test	0,82	0,81	0,83	0,85	0,81	0,85	0,72
RMSE_train	66,2	44,79	38,86	47,58	48,25	49,6	48,78
RMSE_test	103	104,1	101	93,4	106	93,21	127,8
MAPE_train, %	37,1	37,31	37,62	37,27	37,33	37,39	38,56
MAPE_test, %	36,1	36,1	36,42	36	36,15	36	37,29
Время обучения модели, сек	48,9	76,3	32,6	16,2	274	45,75	67,13

Подход №4: TSFresh-агрегация (MinimalFCParametrs) + классификация (RandomForestClassifier) + модель регрессии – гибридный подход

Как и в предыдущих подходах, с помощью методов *TSFresh* (с набором параметров *MinimalFCParameters*) извлекаются признаки из временных рядов данных для каждого трансформатора, данные стандартизируются. На этапе классификации применяется *RandomForestClassifier (*ключевые гиперпараметры: *n_estimators=150* – количество деревьев, соответствующее объему выборки для повышения уровня стабильности и точности предсказаний; ограничение глубины каждого дерева с целью предотвратить переобучение – *max_depth=7.)*, посредством которого на основе извлеченных признаков трансформаторы разделяются на две группы: "нормальные" и "аномальные" с точки зрения прогнозируемого RUL. *RandomForestClassifier* определяет закономерности в данных, позволяющие отделить такие трансформаторы от остальных, присваивая им метку класса (0 или 1)., которая затем используется при построении модели регрессии в тренировочном датасете. Качество классификации оценивается с помощью F1-меры, которая на тестовой выборке составила – 0.79.

Далее происходит обучение модели регрессии, на основе метрик лучшей оказалась модель на базе *CatBoost*. Прогноз классификатора (*RandomForestClassifier*) интегрируются в качестве нового признака к обучающей и тестовой выборкам, после чего модель *CatBoostRegressor* обучается на основе извлеченных признаков TSFresh и меток класса. Из значимых гиперпараметров модели регрессии отметим увеличение значения *l-2* регуляции *l2_leaf_reg=5* с целью предотвращения переобучения модели, а также шаг обновления весов с целью повышения точности модели *learning_rate=0.03*.

Гибридный подход с использованием *RandomForest Classifier* для классификации трансформаторов и *CatBoostRegressor* для регрессии позволил разделить трансформаторы на классы с высокой точностью и построить модель регрессии для прогнозирования RUL с заметным улучшением обобщающей способности модели и снижением переобучения (табл. 3).

Подход №5: TSFresh-агрегация (MinimalFCParametrs) + классификация (RFClassifier) + модель регрессии – гибридный подход.

Для извлечения признаков из временных рядов применяется библиотека TSFresh (с параметрами *EfficientFC Parameters* – в нашем случае 3176 показателей из-за количества временных рядов в исходном наборе данных), извлекается широкий набор признаков, включающий не только базовые статистические характеристики, но и более сложные характеристики, что позволяет модели получить более полную информацию о временных рядах и потенциально повысить точность прогнозирования. Для разделения трансформаторов на "нормальные" и "аномальные" используется модель *RFClassifier*, как и в подходе № 4 (количество деревьев на том же уровне; максимальная величина увеличена до 10; минимальное количество объектов в “leaf” увеличено с целью уменьшения переобучения – *min_samples_leaf=6*, с этой же целью минимальное количество объектов для разбиения узла увеличено *min_samples_split=15;* применяется регуляция; для уменьшения переобучения, количество признаков, используемых для построения каждого дерева ограничено до 50% – *max_features=0.5*). Построение регрессии, инференс модели и оценка качества. На основе полученных метрик (табл. 3) наиболее подходящим методом построения модели регрессии является *ExtraTreesRegressor* (гиперпараметры: *n_estimators=100*; глубина деревьев – *max_depth =10*; *min_samples_split=5* – минимальное число объектов для разбиения узла; *min_samples_leaf=5* – минимальное число объектов в “leaf”).

Значение коэффициента детерминации на тестовой выборке указывает на то, что модель объясняет 85% дисперсии целевой переменной. Значения MAE и RMSE на тестовой выборке относительно невелики, что говорит о хорошей точности предсказаний. Классификатор также работает эффективно, что подтверждается самым высоким значением качества классификации. Сформированный гибрид "TSFresh-агрегация (*MinimalFCParametrs*) / классификация (*RandomForestClassifier*) / модель регрессии (*ExtraTreesRegressor*)" демонстрирует наивысшие показатели качества как на обучающей, так и на тестовой выборке. Графики распределения предсказанных и истинных значений RUL (рис. 2) достаточно близки друг к другу (модель в меру точно аппроксимирует распределение целевых переменных RUL), что подтверждает сформулированные выводы, а смещение прогноза должно быть учтено в ходе дальнейших исследований при оптимизации полученной гибридной модели.

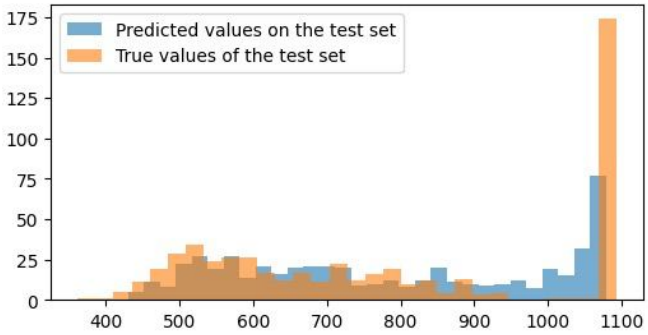

Рис. 2. Сравнение распределений предсказанного и фактического RUL

Заключение

Таким образом, полученную гибридную модель можно использовать в автоматизированной системе прогноза RUL трансформаторов со средней абсолютной ошибкой в 33 дня. Исследование подтверждает, что интеллектуальные системы мониторинга, основанные на методах машинного обучения, необходимы для безопасного и эффективного контроля состояния работоспособности трансформаторов.

Модель может постоянно совершенствоваться за счет расширения выборки, оптимизации гиперпараметров и признаков классификации, за счет применения кластеризации, а также интеграцию с другими системами, что позволяет повышать ее точность и адаптивность к изменяющимся условиям.

Также гибридные модели могут быть дополнены методами самообучения и ансамблевыми алгоритмами с целью получения метамодели [10], что позволит не только улучшать точность прогнозов, но и оперативно адаптироваться к новым эксплуатационным условиям и вызовам.

Литература

1. Bashir T. Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN / T. Bashir, H. Chen, M. Tahir, L. Zhu // Energy Reports. – 2022. – № 8. – С. 1678-1686.

2. Li X. Remaining Useful Life Estimation in Prognostics Using Deep Convolution Neural Networks / X. Li, Q. Ding, J.Q. Sun // Reliability Engineering & System Safety. – 2017. – № 172. – С. 1-11.

3. Moon S. Remanufacturing Decision-Making for Gas Insulated Switchgear with Remaining Useful Life Prediction / S. Moon, H. Cho, E. Koh, Y.S. Cho, H.L. Oh, Y. Kim, S.B. Kim // Sustainability. – 2022. – № 19. – С. 1-13.

4. Ochella S. Adopting machine learning and condition monitoring P-F curves in determining and prioritizing high-value assets for life extension / S. Ochella, M. Shafiee, C. Sansom // Expert Systems with Applications. – 2021. – № 176.

5. Aizpurua J. Adaptive Power Transformer Lifetime Predictions Through Machine Learning and Uncertainty Modeling in Nuclear Power Plants / J. Aizpurua, S. McArthur, B. Stewart, B. Lambert, J. Cross, V. Catterson // IEEE Transactions on Industrial Electronics. – 2019. – № 6. – С. 4726-4737.

6. Chaturvedi A. Performance assessment and RUL prediction of power converters under the multiple components degradation / A. Chaturvedi, M. Sarma, S. Chaturvedi, J. Bernstein // Microelectronics Reliability. – 2023. – № 144.

7. Das S. A Hybrid Regression Model to Estimate Remaining Useful Life of Transformer Liquid / S. Das, A. Paramane, U. Rao, P. Rozga // IEEE Transactions on Dielectrics and Electrical Insulation. – 2024. – № 31. – С. 1062-1069.

8. Bashir T. Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN / T. Bashir, C. Haoyong, M. Tahir, Z. Liqiang // Energy Reports. – 2022. – № 8.

9. Berghout T. Systematic Guide for Predicting Remaining Useful Life with Machine Learning / T. Berghout, M. Benbouzid // Electronics. – 2022. – № 11. – С. 1-31.

10. Katser I. Machine learning methods for anomaly detection in nuclear power plant power transformers / I. Katser, D. Raspopov, V. Kozitsin, M. Mezhov. – URL: https://www.researchgate.net/publication/365633437_Machine_Learning_Methods_for_Anomaly_Detection_in_Nuclear_Power_Plant_Power_Transformers (дата обращения 14.02.2025).

11. Кацер Ю. Все, что вы хотели знать о задаче определения остаточного ресурса оборудования. – URL: https://habr.com/ru/articles/717812/ (дата обращения 14.02.2025).

12. Katser I. Power Transformers FDD and RUL [Data set]. – URL: https://www.kaggle.com/datasets/yuriykatser/power-transformers-fdd-and-rul/data (дата обращения 06.02.2025).

13. Ring M. A Survey of Network-based Intrusion Detection Data Sets // Comput. Secur. – 2019. – № 86. – С. 147-167.

14. Christ M. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package) / M. Christ, N. Braun, J. Neuffer, A. W. Kempa-Liehr // Neurocomputing. – 2018. – № 307. – С. 72-77.

15. https://tsfresh.readthedocs.io/en/latest/index.html (дата обращения 12.02.2025).

16. Bentéjac C. A comparative analysis of gradient boosting algorithms / C. Bentéjac, A. Csörgő, G. Martínez-Muñoz // Artificial Intelligence Review. – 2021. – № 54. – С. 1937-1967.

17. Karthick K. Optimizing Electric Vehicle Battery Life: A Machine Learning Approach for Sustainable Transportation / K. Karthick, S. Ravivarman, R. Priyanka // World Electric Vehicle Journal. – 2024. – № 15.

18. https://xgboost.readthedocs.io/en/latest/#xgboost-documentation (дата обращения 11.02.2025).

19. https://lightgbm.readthedocs.io/en/stable/ (дата обращения 13.02.2025).
20. https://scikit-learn.org/stable/api/sklearn.html (дата обращения 14.02.2025).
21. Purushotam N. Crop Yield Prediction Using Gradient Boosting Neural Network Regression Model / N. Purushotam, R. V. Lakshmana, C. Gunturu, A. Niharika, C. Anupama, G. Srivalli // International Journal on Recent and Innovation Trends in Computing and Communication. – 2023. – № 11.
22. Кацер Ю. Решение задачи определения RUL трансформаторов с помощью машинного обучения на python. – URL: https://habr.com/ru/articles/743682/ (дата обращения 08.02.2025).

DOI: 10.24892/RIJIE/20250302

Machine Learning Methods for Predicting the Remaining Useful Life of Equipment: a Hybrid Approach

Alchakov V.V., Pisarev V.I.
Sevastopol State University
Sevastopol, Russian Federation
alchakov@sevsu.ru, pisarev.valsidalv@gmail.com

Abstract. **This paper addresses the problem of predicting the remaining useful life (RUL) of power transformers based on the analysis of dissolved gas concentrations in transformer oil. The study is based on time-series measurements of H_2, CO, C_2H_2, and C_2H_2 concentrations over an extended period. The dataset was preprocessed, aggregated, and analysed to identify patterns and factors influencing equipment degradation. Machine learning techniques were employed for RUL prediction, including gradient boosting models (XGBoost, LightGBM, CatBoost), decision trees (RandomForest, ExtraTrees), and multilayer perceptrons (MLP). Various data aggregation strategies were considered, including conventional averaging and automated feature extraction using TSFresh. Additionally, feature classification was performed to enhance the accuracy of regression-based predictions. The study resulted in a hybrid approach that integrates classification and regression, effectively improving predictive performance in estimating the remaining useful life of the analysed equipment.**

Keywords: **machine learning, hybrid model, remaining useful life, regression, predictive model.**

Библиографическое описание статьи

Альчаков В.В. Методы машинного обучения в задачах прогнозирования остаточного ресурса оборудования: гибридный подход / В.В. Альчаков, В.И. Писарев // Машиностроение: сетевой электронный научный журнал. – 2025. – Т.12, №3. – С. 9-16. DOI: 10.24892/RIJIE/20250302

Reference to article

Alchakov V.V., Pisarev V.I. Machine learning methods for predicting the remaining useful life of equipment: a hybrid approach, *Russian Internet Journal of Industrial Engineering*, 2025, vol.12, no.3, pp. 9-16. DOI: 10.24892/RIJIE/20250301
