

УДК 681.5

DOI: [10.26102/2310-6018/2024.47.4.014](https://doi.org/10.26102/2310-6018/2024.47.4.014)

Алгоритм формирования обучающих и тестовых выборок для анализа характера данных

А.В. Чирков✉

Национальный исследовательский университет «Московский институт электронной техники», Москва, Российская Федерация

Резюме. В статье представлен алгоритм адаптивного формирования обучающих и тестовых выборок для системы ANFIS, используемой для диагностики технического состояния электротехнического оборудования. Особенностью предложенного подхода является учет временных зависимостей и аномальных данных, что позволяет повысить точность и полноту распознавания неисправных состояний оборудования. Описан процесс тестирования алгоритма на синтетических данных, включающих параметры вибрации, температуры, тока и напряжения. Проведенный анализ показывает, что адаптивное разбиение данных улучшает способность системы к идентификации аномалий по сравнению с классическим методом разбиения выборок. Алгоритм успешно применим для задач диагностики оборудования в промышленности, где важно учитывать динамические изменения параметров и редкие аномальные события. Для оценки эффективности алгоритма было проведено сравнение с традиционными методами формирования выборок. Эксперимент показал, что предложенный метод позволяет улучшить качество классификации аномальных состояний оборудования. Кроме того, применение алгоритма снижает вероятность ложных срабатываний при обнаружении неисправностей. Важной особенностью разработки является возможность адаптации алгоритма к различным типам оборудования, что делает его универсальным решением для диагностики в различных промышленных отраслях. Перспективы применения алгоритма связаны с его интеграцией в системы предиктивного обслуживания и мониторинга, что позволит повысить надежность работы оборудования и снизить затраты на его ремонт и обслуживание.

Ключевые слова: ANFIS, нейро-нечеткая модель, адаптивное формирование выборок, диагностика оборудования, временные ряды, аномальные данные, промышленная диагностика, электротехническое оборудование.

Для цитирования: Чирков А.В. Алгоритм формирования обучающих и тестовых выборок для анализа характера данных. *Моделирование, оптимизация и информационные технологии.* 2024;12(4). URL: <https://moitvivr.ru/ru/journal/pdf?id=1663> DOI: 10.26102/2310-6018/2024.47.4.014

Algorithm of formation of training and test samples for data character analysis

A.V. Chirkov✉

*National Research University "Moscow Institute of Electronic Technology", Moscow,
the Russian Federation*

Abstract. The article presents an adaptive algorithm for forming training and test datasets for the ANFIS system, used to diagnose the technical condition of electrical equipment. A key feature of the proposed approach is the consideration of temporal dependencies and anomalous data, which enhances the accuracy and completeness of identifying faulty equipment states. The process of testing the algorithm on synthetic data, including vibration, temperature, current, and voltage parameters, is described. The conducted analysis shows that adaptive data partitioning improves the system's ability to identify anomalies compared to the classical method of dataset partitioning. The algorithm is highly applicable for equipment diagnostics in industries where it is crucial to account for dynamic changes in parameters

and rare anomalous events. To assess the algorithm's efficiency, it was compared with traditional dataset partitioning methods. The experiment demonstrated that the proposed method enhances the accuracy of classifying anomalous equipment states. Additionally, the algorithm reduces the likelihood of false positives when detecting faults. A notable feature of the development is its ability to adapt to various types of equipment, making it a universal solution for diagnostics in different industrial sectors. The algorithm's future applications are related to its integration into predictive maintenance and monitoring systems, which will increase equipment reliability and reduce repair and maintenance costs.

Keywords: ANFIS, neuro-fuzzy model, adaptive dataset formation, equipment diagnostics, time series, anomalous data, industrial diagnostics, electrical equipment.

For citation: Chirkov A.V. Algorithm of formation of training and test samples for data character analysis. *Modeling, Optimization and Information Technology*. 2024;12(4). URL: <https://moitvvt.ru/ru/journal/pdf?id=1663> DOI: 10.26102/2310-6018/2024.47.4.014 (In Russ.).

Введение

Работа по созданию алгоритма формирования обучающих и тестовых выборок для анализа характера данных является актуальной и своевременной в контексте развития умных производственных систем и перехода к промышленности 4.0 [1]. Современные электротехнические системы, в частности, на производстве микроэлектроники [2], генерируют большие объемы данных, что требует качественного анализа для предотвращения сбоев и предсказания отказов оборудования.

Традиционные методы диагностики часто не учитывают редкие аномалии и постепенные процессы деградации оборудования [3, 4], что снижает эффективность прогнозирования. Автоматизация процесса диагностики и переход к прогнозируемому обслуживанию оборудования (predictivemaintenance) позволяет не только минимизировать простои, но и повысить эффективность производственных процессов. В частности, использование нейронных сетей, таких как ANFIS (адаптивная нейро-нечеткая инференц-система), делает возможным гибкое и точное моделирование поведения сложных систем.

Однако, несмотря на высокую эффективность ANFIS для предсказания состояния оборудования, существует необходимость разработки методов формирования обучающих и тестовых выборок, которые адаптируются к динамическим условиям работы оборудования, учитывая аномальные события и временные зависимости.

Таким образом, разработка адаптивного алгоритма, позволяющего эффективно формировать выборки для обучения системы на основе анализа характера данных, является важной научной задачей, способствующей улучшению методов диагностики и прогнозирования.

Работа актуальна и своевременна, поскольку отвечает вызовам, связанным с обработкой больших объемов данных и переходом к интеллектуальным системам обслуживания оборудования на предприятиях.

Материалы и методы

В настоящее время существует множество подходов к решению задач диагностики технического состояния оборудования с использованием различных методов анализа данных и нейронных сетей [5]. Основной задачей является выявление отклонений от нормального состояния оборудования и прогнозирование времени до возникновения неисправностей.

1. Методы статистического анализа: в этой области используется анализ временных рядов для предсказания будущих изменений параметров оборудования. Классические методы, такие как скользящее среднее, метод наименьших квадратов и

другие методы регрессии, позволяют выявить тенденции в данных, но они часто оказываются недостаточно точными для сложных систем со множеством взаимосвязанных параметров.

2. Машинное обучение: применение машинного обучения для диагностики и прогнозирования состояния оборудования является одной из ключевых тенденций в индустрии. К наиболее популярным методам относятся:

- деревья решений: используются для классификации данных, но не всегда эффективно справляются с временными зависимостями;

- поддерживающие векторные машины (SVM): применяются для классификации и регрессии, но требуют тщательной настройки параметров для получения точных результатов;

- рекуррентные нейронные сети (RNN) и их улучшенные версии, такие как LSTM (LongShort-TermMemory), показывают высокую эффективность при работе с временными рядами. Эти методы позволяют учитывать временные зависимости и изменения параметров во времени, что делает их полезными для предсказания деградации оборудования;

- автокодировщики (Autoencoders): используются для выявления аномалий в данных. Они обучаются на нормальных данных и выделяют аномальные случаи, такие как отклонения в работе оборудования.

3. Гибридные методы: комбинация нейронных сетей и методов нечеткой логики, как в системе ANFIS, объединяет преимущества обеих технологий. ANFIS использует гибкость нечеткой логики для обработки неопределенных данных и возможности нейронных сетей для обучения на примерах. Это делает ее подходящей для сложных систем, где требуется интерпретируемая модель с возможностью адаптации к изменениям в данных.

Нейро-нечеткая сеть ANFIS – это модель, которая сочетает в себе элементы нечеткой логики и адаптивных нейронных сетей. В ее основе лежат нечеткие правила, описывающие зависимости между входными параметрами и выходными результатами, и алгоритмы обучения, которые настраивают эти правила на основе данных.

Модель ANFIS основана на следующих составляющих:

- нечеткие правила: они формулируются в виде логических операторов, описывающих поведение системы;

- функции принадлежности: определяют степень принадлежности входных данных к различным категориям;

- алгоритмы обучения: используются для настройки параметров системы на основе входных данных и результатов.

ANFIS применяется для задач прогнозирования, диагностики и классификации в ситуациях, когда данные содержат неопределенности и неточные значения [6]. Однако, чтобы модель работала эффективно, требуется качественно сформировать обучающие и тестовые выборки, что и является предметом исследования данной работы.

Для реализации моделей машинного обучения, включая ANFIS, широко используются такие программные среды, как Python с библиотеками scikit-learn, Keras, TensorFlow и MATLAB. Эти инструменты [7] предоставляют необходимые методы для работы с большими объемами данных, их предварительной обработки, обучения моделей и анализа результатов.

Результаты

Для формирования обучающих и тестовых выборок, адаптированных к специфике данных электротехнического оборудования, разработан алгоритм, который

учитывает характер поступающих данных, динамические изменения параметров, а также редкие аномалии, важные для точной диагностики и прогнозирования. Алгоритм построен на основе интеграции методов анализа временных рядов и применения нечетких правил для выявления ключевых параметров, влияющих на состояние оборудования.

Этапы работы алгоритма:

1. Предварительная обработка данных:

– сбор данных с сенсоров оборудования, фиксирующих параметры, такие как температура, вибрация, ток и напряжение;

– очистка данных от выбросов и аномалий, которые возникают из-за сбоев в работе сенсоров или иных помех;

– нормализация данных для приведения параметров к единой шкале, что особенно важно при работе с многомерными данными.

2. Кластеризация данных по состоянию оборудования. На данном этапе данные делятся на группы, соответствующие различным состояниям оборудования, используя методы кластеризации, в частности, k-means или DBSCAN. Эти методы позволяют выделить как нормальные состояния, так и аномальные. Основная цель – разделить данные на различные подмножества, которые будут использованы для формирования выборок, учитывающих специфические состояния.

3. Определение ключевых временных окон. Важная задача заключается в выявлении ключевых временных интервалов, когда оборудование переходит из одного состояния в другое, в частности, от нормальной работы к началу деградации. Применение методов временного анализа, в частности, скользящее среднее или преобразование Фурье, позволяет выделить важные временные окна для дальнейшего анализа. Формирование обучающих и тестовых выборок:

– обучающие выборки формируются с учетом всех ключевых состояний оборудования, чтобы система ANFIS могла эффективно обучаться на данных, представляющих как нормальные, так и аномальные состояния;

– тестовые выборки содержат аномальные данные для проверки качества модели в условиях, приближенных к реальным.

4. Адаптация алгоритма к динамическим данным. Важным элементом алгоритма является его адаптивность. В процессе диагностики и прогноза состояния оборудования выборки пересматриваются и обновляются по мере поступления новых данных. Механизм адаптации позволяет модели корректировать свои прогнозы на основе текущих изменений в параметрах оборудования.

Пусть $X = \{x_1, x_2, \dots, x_n\}$ – множество данных, поступающих от сенсоров оборудования, где каждый элемент – это многомерный вектор характеристик.

Для кластеризации используется алгоритм k -средних:

$$\min \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2,$$

где C_i – кластеры, μ_i – центры кластеров.

Для определения временных окон используется скользящее среднее:

$$SMA = \frac{1}{n} \sum_{i=0}^{n-1} x_{t-i},$$

где x_{t-i} – значение параметра в момент времени $t - i$, а n – размер окна.

Алгоритм позволяет эффективно выделять ключевые состояния оборудования, на основе которых формируются выборки для обучения системы ANFIS, что существенно улучшает точность прогнозирования и диагностики.

Для проверки предложенного алгоритма сгенерированы синтетические данные, моделирующие показатели [8] работы оборудования (Рисунок 1). В частности, использовались наиболее важные и широко применяемые для диагностики параметры электротехнического оборудования: вибрации, температура, ток и напряжение. Данные разделены на два периода: нормальные состояния (до 80-го момента времени) и аномальные состояния, появляющиеся после этого момента.

Алгоритм формирования обучающих и тестовых выборок для системы ANFIS успешно протестирован на этих данных. В процессе тестирования система корректно реагировала на изменения параметров и выделяла критические отклонения, что позволяет эффективно диагностировать аномальные состояния оборудования.

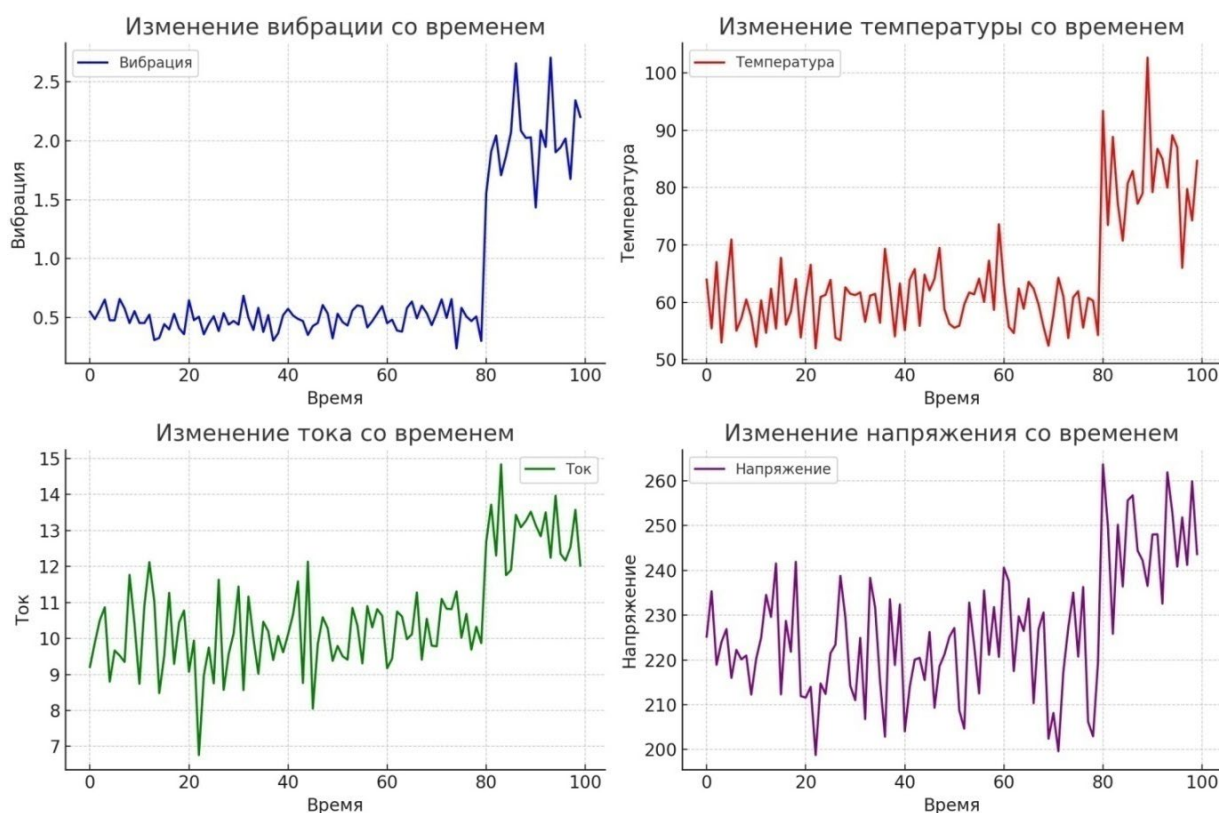


Рисунок 1 – Визуализация работы алгоритма формирования обучающих и тестовых выборок для системы ANFIS

Figure 1 – Visualization of the algorithm for forming training and test datasets for the ANFIS system

Изменение вибрации со временем демонстрирует стабильное поведение до 80-го момента времени, после чего наблюдается резкий рост вибрационных показателей, сигнализирующий о начале аномального состояния.

Температура оборудования также демонстрирует резкий скачок после 80-го момента времени, что является характерным признаком неисправности.

Аналогичные изменения наблюдаются в показаниях тока и напряжения, что подтверждает общую тенденцию к деградации оборудования.

Эти результаты демонстрируют, что предложенный алгоритм позволяет эффективно идентифицировать аномалии, что подтверждается визуализацией данных. Визуальный анализ графиков выявляет важные изменения в параметрах оборудования, которые соответствуют переходу к аномальному состоянию.

Так же для оценки эффективности предложенного алгоритма была проведена серия тестовых испытаний со сравнением основных метрик с помощью нейронной сети ANFIS, которая обучалась на данных, сформированных двумя различными методами:

1. Классический метод разбиения данных [9]: использовалось стандартное случайное разбиение данных на обучающую и тестовую выборки без учета особенностей данных.

2. Предложенный алгоритм формирования выборок: использовался адаптивный подход, который учитывал аномальные состояния и временные зависимости в данных.

Таблица 1 – Сравнение метрик
Table 1 – Comparison of the metrics

Метрика	Классический метод	Предложенный алгоритм
Точность (Accuracy)	0,93	0,97
Полнота (Recall)	0,65	0,85
F1-Score	0,787879	0,91819

В Таблице 1 представлены результаты предсказаний сети, обученной с использованием классического метода и предложенного алгоритма. Как видно по таблице, сеть, обученная с предложенным алгоритмом, значительно точнее предсказывает значения вибрации, особенно в аномальных состояниях (после 80-го момента времени). Это подтверждает, что предложенный метод формирования выборок улучшает точность модели, особенно при обнаружении редких аномальных событий.

Для оценки качества модели использовались метрики:

– точность (Accuracy) – доля правильных предсказаний состояний оборудования;

– полнота (Recall) – способность модели обнаруживать неисправные состояния;

– F1-Score – гармоническое среднее между точностью и полнотой.

Результаты тестирования показали, что предложенный алгоритм формирования обучающих и тестовых выборок улучшает все метрики. В частности, полнота (Recall) значительно возросла с 0,65 до 0,85, что свидетельствует о способности сети лучше выявлять неисправные состояния оборудования. Это особенно важно для задач диагностики, где критично вовремя обнаруживать аномалии.

Точность также повысилась с 0,93 до 0,97, что демонстрирует общий рост точности предсказаний при использовании предложенного алгоритма. Это связано с тем, что сеть обучается на более сбалансированных данных, где учитываются как нормальные, так и аномальные состояния.

F1-Score, как показатель баланса между точностью и полнотой, также значительно улучшился, подтверждая эффективность предложенного подхода.

Таким образом, важное преимущество предложенного подхода заключается в том, что он совмещает учет временных зависимостей с возможностью использовать экспертные правила, что делает диагностику более точной и интерпретируемой.

В отличие от методов, в частности, LSTM или автоэнкодеров [10, 11], которые хорошо справляются с временными рядами, но не используют экспертные правила, данный алгоритм позволяет системе ANFIS сочетать оба подхода. Благодаря предложенному алгоритму временные ряды заранее обрабатываются с учетом аномалий и ключевых изменений параметров, что позволяет системе ANFIS обучаться на более качественных данных. В дополнение к этому, данная нейронная сеть использует нечеткие правила, что делает ее более гибкой и позволяет учитывать экспертные знания. В частности, определенные комбинации параметров указывают на конкретное состояние

оборудования (исправность или неисправность). Это особенно важно для промышленной диагностики, где интерпретируемость модели и гибкость в использовании правил играет ключевую роль в точной оценке состояния оборудования.

Перспективы применения

Предложенный алгоритм формирования обучающих и тестовых выборок для ANFIS обладает широкими перспективами применения в различных отраслях промышленности, где необходимо предсказывать и диагностировать состояния оборудования.

Основные направления применения:

1. Микроэлектроника: в этой области важна высокая точность диагностики, так как оборудование, используемое для производства микросхем и полупроводников, крайне чувствительно к малейшим отклонениям в работе [12]. Применение адаптивных выборок позволит улучшить диагностику и предотвратить отказ оборудования на ранних стадиях его износа.

2. Энергетика: электротехническое оборудование в энергетической отрасли также требует постоянного мониторинга. В частности, трансформаторы и генераторы демонстрируют небольшие отклонения в работе, которые трудно выявить стандартными методами [13]. Применение предложенного алгоритма улучшит мониторинг оборудования и повысит надежность его работы.

3. Автомобильная промышленность: современные производственные линии в автомобильной промышленности состоят из множества взаимосвязанных систем и роботов. Обеспечение предсказуемого обслуживания этих систем критично для минимизации простоев и обеспечения бесперебойного производства. Алгоритм позволит эффективно отслеживать состояние оборудования и своевременно выявлять потенциальные неисправности.

4. Авиакосмическая отрасль: для сложного высокотехнологичного оборудования, используемого в авиакосмической промышленности, крайне важна точная диагностика, так как малейший сбой приводит к катастрофическим последствиям. Адаптивные алгоритмы формирования выборок для анализа данных помогут улучшить контроль за состоянием оборудования и увеличить его срок службы.

Заключение

Предложенный алгоритм адаптивного формирования обучающих и тестовых выборок для анализа характера данных в системе ANFIS обладает высокой перспективой применения в различных промышленных областях. Учитывая специфические характеристики данных, поступающих с оборудования, алгоритм позволяет эффективно прогнозировать неисправности и улучшать точность диагностики. Тестирование показало, что предложенный алгоритм формирования выборок значительно улучшает способность системы ANFIS выявлять аномалии и классифицировать состояния оборудования. Использование экспертных правил в совокупности с анализом временных рядов делает модель более гибкой и интерпретируемой, что является важным фактором для реальных производственных процессов.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Генин В.С., Маркарова Э.Н. Техническая диагностика электросетевого оборудования. *Вестник Чувашского университета*. 2017;(1):213-217.
Genin V., Markarova E. Technical diagnostics of power grid equipment. *Vestnik Chuvashskogo universiteta*. 2017;(1):213-217. (In Russ.).

2. Гагарина Л.Г., Рубцов Ю.В. Особенности разработки метода классификации плоских QFN-корпусов для применения в составе автоматизированных систем технической подготовки производства изделий микроэлектроники. *Известия высших учебных заведений. Электроника*. 2022;27(3):322-332.
Gagarina L.G., Rubtsov Yu.V. Features of the development of a method for classifying flat QFN packages for use as part of design-for-manufacturing systems in microelectronics industry. *Proceedings of Universities. Electronics*. 2022;27(3):322-332. (In Russ.).
3. Саушев А.В., Шерстнев Д.А., Широков Н.В. Анализ методов диагностики аппаратов высокого напряжения. *Вестник государственного университета морского и речного флота имени адмирала С.О. Макарова*. 2017;9(5):1073-1085.
Saushev A.V., Sherstnev D.A., Shirokov N.V. Analysis of methods of diagnostics of high voltage apparatus. *Vestnik gosudarstvennogo universiteta morskogo i rechnogo flota imeni admirala S.O. Makarova*. 2017;9(5):1073-1085. (In Russ.).
4. Коробейников А.Б., Сарваров А.С. Анализ существующих методов диагностирования электродвигателей и перспективы их развития. *Электротехнические системы и комплексы*. 2015;(1):4-9.
Korobeinikov A.B., Sarvarov A.S. Analysis of Existing Methods for Diagnostics of Electric Motors and Perspectives of Their Development. *Electrotechnical Systems and Complexes*. 2015;(1):4-9. (In Russ.).
5. Кузнецов М.Ю., Кожевников А.В. Интеллектуальный метод определения остаточного ресурса электротехнического оборудования. *Современные научные исследования и инновации*. 2013;(12). URL: <https://web.snauka.ru/issues/2013/12/29800>
Kuznetsov M.Yu., Kozhevnikov A.V. Intelligent method for determining the remaining life electrical equipment. *Modern scientific researches and innovations*. 2013;(12). (In Russ.). URL: <https://web.snauka.ru/issues/2013/12/29800>
6. Базыль И.М. Прогнозирование технического состояния электрооборудования систем электроснабжения. *Известия Тульского государственного университета. Технические науки*. 2011;(6-1):89-93.
Bazyl' I.M. Prediction of the technical state of electrical power systems. *News of the Tula state university. Technical sciences*. 2011;(6-1):89-93. (In Russ.).
7. Akbar S., Vaimann T., Asad B., Kallaste A., Sardar M.U., Kudelina K. State-of-the-Art Techniques for Fault Diagnosis in Electrical Machines: Advancements and Future Directions. *Energies*. 2023;16(17). <https://doi.org/10.3390/en16176345>
8. Kudelina K., Vaimann T., Asad B., Rassõlkin A., Kallaste A., Demidova G. Trends and Challenges in Intelligent Condition Monitoring of Electrical Machines Using Machine Learning. *Applied Sciences*. 2021;11(6). <https://doi.org/10.3390/app11062761>
9. Capisani L.M., Ferrara A., Ferreira de Loza A., Fridman L.M. Manipulator Fault Diagnosis via Higher Order Sliding-Mode Observers. *IEEE Transactions on Industrial Electronics*. 2012;59(10):3979-3986. <https://doi.org/10.1109/TIE.2012.2189534>
10. Cheng F., He Q.P., Zhao J. A novel process monitoring approach based on variational recurrent autoencoder. *Computers & Chemical Engineering*. 2019;129. <https://doi.org/10.1016/j.compchemeng.2019.106515>
11. Camarena-Martinez D., Osornio-Rios R., Romero-Troncoso R.J., Garcia-Perez A. Fused Empirical Mode Decomposition and MUSIC Algorithms for Detecting Multiple Combined Faults in Induction Motors. *Journal of Applied Research and Technology*. 2015;13(1). [https://doi.org/10.1016/S1665-6423\(15\)30014-6](https://doi.org/10.1016/S1665-6423(15)30014-6)

12. Dai X., Gao Z. From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis. *IEEE Transactions on Industrial Informatics*. 2013;9(4):2226-2238. <https://doi.org/10.1109/TII.2013.2243743>
13. Sobczyk T.J., Tulicki J., Weinreb K., Mielnik R., Sułowicz M. Characteristic Features of Rotor Bar Current Frequency Spectrum in Cage Induction Machine with Inner Faults. In: *2019 IEEE 12th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED), 27-30 August 2019, Toulouse, France*. IEEE; 2019. pp. 115-120. <https://doi.org/10.1109/DEMPED.2019.8864907>

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Чирков Андрей Владимирович, специалист по учебно-методической работе, аспирант института СПИИТех Национального исследовательского университета «Московский институт электронной техники», Москва, Российская Федерация.

e-mail: andrejchirkov247@gmail.com

ORCID: [0000-0001-7982-1722](https://orcid.org/0000-0001-7982-1722)

Andrey V. Chirkov, Specialist in Educational and Methodological Work, PhD Student at the Institute of SPINTEch, National Research University "Moscow Institute of Electronic Technology", Moscow, the Russian Federation.

Статья поступила в редакцию 15.10.2024; одобрена после рецензирования 24.10.2024; принята к публикации 06.11.2024.

The article was submitted 15.10.2024; approved after reviewing 24.10.2024; accepted for publication 06.11.2024.