На правах рукописи

БАХТИН ВАДИМ ВЯЧЕСЛАВОВИЧ

МЕТОД СИНТЕЗА НЕЙРОСЕТЕВЫХ УСТРОЙСТВ ДЛЯ РЕАЛИЗАЦИИ РЕЖИМА FOG COMPUTING

2.3.2 – Вычислительные системы и их элементы

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Пермь 2023

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Пермский национальный исследовательский политехнический университет»

Научный руководитель	**ТЮРИН Сергей Феофентович,** доктор технических наук, профессор
Официальные оппоненты	**ВИНОГРАДОВ Геннадий Павлович,** доктор технических наук, доцент, ФГБОУ ВО «Тверской государственный технический университет», кафедра «Информатика и прикладная математика», профессор
	СОКОЛОВА Юлия Васильевна, кандидат технических наук, АО «Научно-производственное объединение им. С.А. Лавочкина», ведущий специалист
Ведущее предприятие	Федеральное государственное бюджетное образовательное учреждение высшего образования «Юго-Западный государственный университет», г. Курск

Защита состоится «30» июня 2023 г. в 16.00 часов на заседании диссертационного совета Пермского национального исследовательского политехнического университета Д ПНИПУ.05.14, по адресу: 614990, г. Пермь, Комсомольский проспект, д. 29, ауд. 345.

С диссертацией можно ознакомиться в библиотеке и на сайте ФГАОУ ВО «Пермский национальный исследовательский политехнический университет» (http://pstu.ru).

Автореферат разослан «5» мая 2023 г.

Ученый секретарь
диссертационного совета Д ПНИПУ.05.14,
доктор технических наук, доцент В.И. Фрейман

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. На первых этапах развития вычислительных систем обработка данных производилась на одном и том же вычислительном устройстве. По мере развития микропроцессорной техники и сетевых технологий появилась возможность передавать по вычислительной сети крупные массивы данных за короткое время, после этого появляется концепция распределенных вычислений. Туманные вычисления – это метод распределения вычислительных задач в виде небольших блоков на небольшие устройства, которые обрабатывают информацию в процессе ее передачи от отправителя к получателю. Возможно использование для реализации туманных вычислений не только персональных компьютеров, смарт-часов, смартфонов и других гаджетов, но и более простых устройств, содержащих недорогие аппаратные средства.

В настоящее время активно развиваются нейросетевые технологии и их реализация в рамках вычислительных систем и их элементов. Необходимость развития нейросетевых технологий в области медицинской диагностики резко возросла с началом пандемии короновируса. Дальнейшие события привели к возрастанию потребностей совершенствования интеллектуальных видов вооружения и военной техники. Нейронная сеть – последовательность из нескольких слоев математических нейронов, соединенных друг с другом. Нейронную сеть, все слои которой производят свои вычисления на одном и том же вычислительном устройстве, принято называть **монолитной нейронной сетью (МНС)**. Нейронную сеть, разбитую на набор связанных блоков, принято называть **блочной нейронной сетью (БНС)**.

В настоящее время возрастает значимость оптимизации использования существующих вычислительных систем. Для достижения технологического суверенитета потребуется решить множество вычислительных задач, в том числе осуществлять нейросетевые вычисления. Расширение обработки данных вычислительной системой зачастую требуется в реальных эксплуатационных задачах. При этом для увеличения возможностей обработки данных вычислительной системой необходимо приобретение и интеграция в систему дополнительных вычислительных узлов, что потребует выделения дополнительных материальных ресурсов. В то же время встречаются ситуации, когда вычислительные мощности уже имеющихся устройств задействованы не в полном объеме. Как следствие, возникает противоречие в практике – с одной стороны, тратятся ресурсы на закупку нового оборудования, с другой стороны, имеются не полностью загруженные мощности в рамках вычислительной системы.

Поэтому целесообразно использование метода синтеза нейросетевых устройств для реализации распределенных нейронных сетей, который мог бы обеспечить оптимальное задействование имеющихся вычислительных устройств таким образом, чтобы не потребовалось устанавливать в существующие системы дополнительных физических вычислителей. При этом необходимо подобрать параметры декомпозиции таким образом, чтобы дополнительная нагрузка в виде нейросетевых вычислений минимально отразилась на исходных параметрах вычислительной системы.

Степень разработанности темы исследования. Началом развития нейронных сетей (НС) стала работа F. Rosenblatt, в которой впервые была предло-

жена нейросетевая модель – персептрон. Большой вклад в развитие НС внесли. J. Redmon, A. Farhadi, А.С. Потапов, М.С. Бурцев (МФТИ), А.А. Южаков, С.Н. Костарев, В.И. Мильке (МГТУ), Ю.Н. Хижняков, Л.Н. Ясницкий, J. Redmon и A. Farhadi предложили архитектуру YOLO для осуществления однопроходной классификации объектов на видео в реальном времени, отечественные ученые также предлагали различные оригинальные архитектуры НС. Развитие архитектур НС и вычислительных сетей привело к созданию класса распределенных НС. Проблемы создания **распределенных НС** в настоящее время рассматриваются в работах таких авторов, как Bradley McDanel, Surat Teerapittayanon, H.T. Kung, Yuchen Liang, W.D. Li, Zhi-Cong Chen, С.Д. Ионов, Ю.В. Алексеенко и др. Были предложены различные методы реализации нейросетевых вычислений в облаке, туманных узлах, на конечных устройствах пользователей. Проблемы **синтеза устройств** рассмотрены в работах Ю.Г. Дьяченко, В.Я. Володарского, Д.А. Гончарова, А.Н. Каменских, С.Ф. Тюрина, Ю.А. Степченкова и др. В их работах были продвижения в решении задачи блочного синтеза устройств, заключающиеся в декомпозиции алгоритма на типовые дискретные устройства (блоки). Однако задачу блочного синтеза устройств нельзя считать до конца решенной, в том числе и в вопросах синтеза нейросетевых устройств.

Объектом исследования является вычислительная система, её элементы и устройства, используемые для реализации искусственных нейронных сетей, ориентированных на туманные вычисления.

Предметом исследования является научно-методический аппарат синтеза вычислительных систем и их элементов, используемых для реализации искусственных нейронных сетей, ориентированных на туманные вычисления.

Цель исследования – решение научной задачи разработки метода синтеза устройств реализации искусственных нейронных сетей, ориентированных на туманные вычисления.

Для достижения поставленной цели в диссертационной работе поставлены и решены следующие **задачи исследования:**

1. Аналитический обзор и анализ моделей, методов и алгоритмов декомпозиции искусственных нейронных сетей в вычислительных системах различных конфигураций и технологий, анализ их недостатков, обоснование актуальности проводимых исследований.
2. Создание математической модели искусственной нейронной сети для синтеза нейросетевых устройств, ориентированных на туманные вычисления.
3. Разработка усовершенствованного метода синтеза устройств реализации искусственных нейронных сетей, ориентированных на туманные вычисления, и его модификации, позволяющей обеспечить отказоустойчивость.
4. Разработка алгоритма преобразования классической нейронной сети в нейронную сеть, адаптированную для туманных вычислений в устройствах.
5. Разработка алгоритма выбора оптимального варианта декомпозиции нейронной сети для реализации на распределенных вычислительных устройствах.
6. Апробация разработанных модели, метода и алгоритмов, реализованных в структуре аппаратного и программного обеспечения элементов вычислительных систем, реализующих распределенные нейронные сети.

Научная новизна заключается в разработанных модели, методе и алгоритмах. Новизна научных результатов диссертационного исследования состоит в том, что:

1. Разработана *математическая модель искусственной нейронной сети для синтеза нейросетевых устройств, ориентированных на туманные вычисления*. Она *отличается* от существующих тем, что с ее помощью возможно балансировать размеры декомпозированных блоков нейронной сети в зависимости от характеристик физических устройств, входящих в вычислительный каскад. Это *позволяет* учитывать требуемую загрузку вычислительных узлов при распределении блоков нейронной сети между различными устройствами.

2. Разработан *метод синтеза устройств реализации искусственных нейронных сетей, ориентированных на туманные вычисления*. Он *отличается* от существующих тем, что учитывает параметры: мощность устройств, оптимальный объем передаваемых между устройствами данных, возможность учета пропорциональности блоков нейронной сети по слоям или по нейронам, а также имеет возможность реализации диагностики и реконфигурации. Это *позволяет* значительно снизить нагрузку на те вычислительные узлы, которые требуется разгрузить в рамках задачи и продолжать работу даже в случае отказа или сбоя части устройств в каскаде.

3. Разработан *алгоритм декомпозиции монолитной нейронной сети на каскад блоков блочной нейронной сети*, *адаптированной для туманных вычислений*. Он *отличается* от существующих тем, что предлагает способ унификации хранения в памяти монолитной нейронной сети и блоков блочной нейронной сети. Это *позволяет* проводить многократную декомпозицию в глубину, например, если потребуется декомпозировать отдельный блок еще на несколько блоков.

4. Разработан *алгоритм выбора оптимального варианта декомпозиции нейронной сети для реализации на распределенных вычислительных устройствах*. Он *отличается* от существующих тем, что реализует многокритериальную оптимизацию путем нахождения Парето-оптимальных вариантов. Это *позволяет* находить оптимальную декомпозицию монолитной нейронной сети сразу по нескольким важным для вычислительной системы параметрам.

Теоретическая значимость заключается в создании модели искусственной нейронной сети, усовершенствованного метода синтеза устройств реализации искусственных нейронных сетей, ориентированных на туманные вычисления, и алгоритмов синтеза и работы устройств, реализующих нейронные сети в режиме fog computing, позволяющих повысить качественные и эксплуатационные характеристики вычислительных систем и их элементов. Разработанные методы и алгоритмы применимы в различных областях, требующих автоматизации с применением нейросетевых методов, в том числе возможна адаптация для реализации нейронных сетей, архитектуры которых отличаются от рассмотренных в работе.

Практическая значимость заключается:

1) в том, что предложенный инструментарий в виде модели, метода, алгоритмов реализован и внедрен в составе аппаратного и программного обеспече-

ния элементов вычислительных систем, реализующих распределенные нейронные сети;

2) в том, что предложенный новый метод декомпозиции монолитной нейронной сети используются в разработках компании «Проминформ», что позволило снизить затраты на создание системы биометрической идентификации на 27 % и сократить энергопотребление прототипа системы биометрической идентификации на 12,7 %. Полученные научные и практические результаты используются в учебном процессе кафедры «Автоматика и телемеханика» Пермского национального исследовательского политехнического университета;

3) в возможности построения отказоустойчивого каскада вычислительных устройств с заранее выбранной глубиной адаптации, реализующего блочную распределенную нейронную сеть;

4) в применимости разработанных методов и алгоритмов в нейросетевых устройствах без изменений разработанных архитектур и параметров НС.

Методология и методы исследования. В диссертационном исследовании используются методы и средства анализа информации о существующих нейронных сетях, математического моделирования нейронной сети, схемотехнического моделирования, анализа и оценки временной сложности алгоритмов. Применяемые методы и средства основаны на положениях дискретной математики, теории булевых функций и автоматов, теории вероятности и математической статистики, теории искусственных нейронных сетей, программирования.

На защиту выносятся следующие научные положения:

1. Существующие методы синтеза нейросетевых устройств позволяют проводить декомпозицию нейронных сетей, однако они не учитывают при этом важные показатели: стоимость, энергопотребление, время работы нейронной сети и время отклика устройств по существующим задачам.

2. Разработанная модель, метод и алгоритмы решают задачу выбора оптимального набора нейросетевых устройств для реализации туманных вычислений с учетом указанных показателей, при этом формируется множество Парето, из которого возможно получить вариант по заданным ограничениям.

3. Полученные научные и практические результаты позволяют снизить материальные затраты и энергопотребление, сохраняя показатели быстродействия в допустимых пределах, и рекомендуются к использованию в областях критического применения вычислительных систем.

На защиту выносятся следующие новые научные результаты:

1. Математическая модель распределенной искусственной нейронной сети для синтеза нейросетевых устройств, ориентированных на туманные вычисления (п. 7 паспорта научной специальности).

2. Метод синтеза устройств реализации искусственных нейронных сетей для работы в режиме туманных вычислений и его модификация, позволяющая обеспечить отказоустойчивость (п. 7 паспорта научной специальности).

3. Алгоритм декомпозиции монолитной нейронной сети на каскад блоков нейронной сети, адаптированной для туманных вычислений (п. 7 паспорта научной специальности).

4. Алгоритм выбора оптимального варианта декомпозиции нейронной сети для реализации на распределенных вычислительных устройствах (п. 7 паспорта научной специальности).

5. Результат апробации разработанных модели, метода и алгоритмов, реализованных в структуре аппаратного и программного обеспечения элементов вычислительных систем (п. 4 паспорта научной специальности).

Достоверность и обоснованность результатов подтверждаются соответствием результатов синтеза нейросетевых устройств и схемотехнического моделирования, которые, в свою очередь, совпали с результатами прототипирования. Достоверность также подтверждается результатами апробации и внедрения предложенных в диссертации модели, метода и алгоритмов в реальные вычислительные системы. Полученные результаты не противоречат теоретическим и практическим положениям, известным из научных публикаций отечественных и зарубежных исследователей в рассматриваемой предметной области.

Апробация работы. Основные теоретические и практические результаты работы докладывались на научно-технических конференциях: Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) (2023, 2022, 2021, 2020, 2019), 14th International Scientific-Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE) (2018), IX Международной научной конференции, посвященной 85-летию профессора В.И. Потапова (Омск, 2022), и в других международных и региональных конференциях, всероссийской научно-технической конференции «Автоматизированные системы управления и информационные технологии» (Пермь, 2022).

Работы по теме диссертационного исследования выполнялись в рамках научного проекта при поддержке РФФИ на средства гранта № 20-37-90036 Аспиранты «Метод синтеза устройств нейросетевого распознавания для реализации режима Fog computing».

Публикации. Основные результаты диссертационной работы опубликованы в 20 научных работах, из них пять статей в журналах, входящих в перечень ведущих журналов и изданий, рекомендуемых ВАК, три в изданиях, индексируемых в базах SCOPUS, два свидетельства о регистрации программы для ЭВМ, остальные – в тезисах докладов, материалах конференций и прочих источниках.

Объем и структура работы. Диссертация состоит из введения, пяти глав, заключения, списка литературы из 104 наименований и пяти приложений. Полный объем диссертации составляет 190 страниц, из которых 138 страниц занимает основной текст диссертации, включающий 43 рисунка и 10 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертации, сформулированы цели и задачи диссертационной работы, раскрыта научная новизна и практическая значимость полученных результатов, сформулированы научные положения, выносимые на защиту.

В **первой главе** приведен анализ существующих методов синтеза нейросетевых устройств, реализации распределенных нейронных сетей, сформулированы научные задачи исследования.

Рассмотрены следующие методы реализации распределенных нейронных сетей: метод реализации нейронных сетей на множестве вычислительных узлов в кластерной среде, метод реализации нейронных сетей с использованием предоставляемых свободных мощностей персональных компьютеров в вычислительной сети, метод, не подразумевающий непосредственно прямой декомпозиции нейронной сети, а выносящий ее выполнение в отдельные вычислительные узлы тумана (Fog Node), и другие.

Как видно из анализа предметной области, различные научные коллективы ведут актуальные исследования в данной области, у рассматриваемой проблемы уже существуют различные методы решения, но они позволяют решить поставленную нами задачу не в полной мере. Однако с учетом ограничений, которые выставляются в нашей работе, и необходимых для этого усовершенствований, рассматриваемая задача еще не была решена.

Исходя из представленных данных, было решено предложить усовершенствованный метод синтеза устройств нейросетевого распознавания для реализации режима Fog Computing. Предлагаемым усовершенствованием существующего метода декомпозиции будет расширение гибкости настройки получаемой декомпозированной блочной нейронной сети за счет различных способов декомпозиции, которые определяются входными параметрами декомпозиции, изменение этих параметров будет приводить к получению распределенных нейронных сетей с различными свойствами, а именно: различным объемом проводимых на конкретных устройствах нейросетевых вычислений и объемом данных, передаваемых между устройствами.

Постановка научной задачи исследования: метод должен обеспечивать поиск оптимального по стоимости, энергопотреблению, времени выполнения нейросетевых вычислений и времени выполнения основных вычислений, решения для декомпозиции нейронной сети на каскад распределенных вычислителей при заданных ограничениях на выбранную архитектуру вычислительной системы (число доступных вычислительных устройств n), их вычислительные мощности $\overline{p}$ и пропускной способности каналов связи q.

1. Оценка стоимости $\boldsymbol{C}$:

$$C = \left\{ C_{\tau_1}\left(n, \overline{p}, q\right), C_{\tau_2}\left(n, \overline{p}, q\right), \ldots . C_{\tau_e}\left(n, \overline{p}, q\right) \right\}. \quad (1)$$

2. Оценка времени выполнения нейросетевых вычислений $\boldsymbol{T^1}$:

$$T^1 = \left\{ T^1_{\tau_1}\left(n, \overline{p}, q\right), T^1_{\tau_2}\left(n, \overline{p}, q\right), \ldots, T^1_{\tau_e}\left(n, \overline{p}, q\right) \right\}. \quad (2)$$

3. Оценка среднего времени отклика пульта на входные сигналы $\boldsymbol{T^2}$:

$$T^2 = \left\{ T^2_{\tau_1}\left(n, \overline{p}, q\right), T^2_{\tau_2}\left(n, \overline{p}, q\right), \ldots, T^2_{\tau_e}\left(n, \overline{p}, q\right) \right\}. \quad (3)$$

4. Оценка итогового энергопотребления (мощности) $\boldsymbol{W}$:

$$W = \left\{ W_{\tau_1}\left(n, \overline{p}, q\right), W_{\tau_2}\left(n, \overline{p}, q\right), \ldots W_{\tau_e}\left(n, \overline{p}, q\right) \right\}. \quad (4)$$

При получении оценок необходимо учесть существующие ограничения современных вычислительных устройств. Для подтверждения работоспособности методов следует выполнить схемотехническое и физическое моделирование решений (декомпозиций) $\Omega = \left\{ \omega_1, \omega_2, \ldots \omega_u \right\}$.

Получить оптимальный набор H элементов, используя метод Парето-оптимизации:

$H = \omega_1(n, \overline{p}, q), \omega_2(n, \overline{p}, q), ...\omega_u(n, \overline{p}, q),$

такой, что $C(H) \to \min, T^2(H) \to \min$ без ухудшения $C(H)$,

$T^1(H) \to \min$ без ухудшения $C(H) u T^2(H)$,

$W(H) \to \min$ без ухудшения $C(H), T^2(H)$ и $T^1(H)$.

Необходимо осуществить декомпозицию нейронной сети на несколько устройств с поиском оптимума по параметрам: стоимость, тепловыделение, энергопотребление (при прочих равных приоритет будет отдан варианту, в котором время выполнения будет лучшим).

Находить искомые параметры будем по следующим формулам:

1. Оценка стоимости ***C*:**

$$C = \sum_{i=0}^{n-1} C_i. \quad (5)$$

2. Оценка времени выполнения нейросетевых вычислений T^1:

$$T^1 = \sum_{i=0}^{n-1} T_{ii}^{exec} + \sum_{j=1}^{n-1} T_j^{send}. \quad (6)$$

3. Оценка среднего времени отклика пульта на входные сигналы $\boldsymbol{T^2}$:

$$T^2 = \overline{T_{pult}} + T_{pult}^{exec}. \quad (7)$$

4. Оценка итогового энергопотребления (мощности) $\boldsymbol{W}$:

$$W = \sum_{i=0}^{n-1} W_i. \quad (8)$$

Для этого требуется разработать комплекс алгоритмов для декомпозиции нейронной сети на блоки для работы распределенной нейронной сети и для поиска оптимальной по предложенным оценкам декомпозиции. Для практической реализации требуется разработать программное обеспечение, реализующее эти алгоритмы. Кроме того, необходимо выполнить проверку результатов синтеза с помощью схемотехнического моделирования и прототипирования.

С целью построения активно отказоустойчивых вычислительных устройств нейронных сетей в режиме туманных вычислений с заранее выбранной глубиной адаптации G требуется модифицировать метод синтеза нейросетевых устройств для реализации туманных вычислений.

Во **второй главе** выполнена формулировка математической модели искусственной нейронной сети (НС), ориентированной на туманные вычисления, а также приведено теоретическое обоснование достоверности математической модели и показана адекватность математической модели.

В первую очередь были сформулированы ограничения, в рамках которых будет работать предлагаемая математическая модель. Предлагаемая математиче-

ская модель отображает многослойные нейронные сети. Разработаны математические модели нейронных сетей со архитектурами: FFNN, сверточные и рекуррентные. Для корректной работы метода синтеза устройств требуется, чтобы к моменту декомпозиции нейронные сети уже были обучены.

Дано: монолитная многослойная нейронная сеть X, результат работы которой – последовательность сигналов $\left\{y_0^K,...,y_{H_K}^K\right\}$.

Найти: последовательность нейронных сетей $\left\{\bar{X}_0,...,\bar{X}_{D-1}\right\}$, где результат вычисления $\bar{X}_0 \rightleftarrows \bar{X}_1 \rightleftarrows \cdots \rightleftarrows \bar{X}_{D-1}$ совпадает с результатом работы сети X.

Процесс преобразования монолитной НС в каскад блочных НС, результат вычисления которого совпадает с результатами монолитной НС, принято называть **декомпозицией** искусственной нейронной сети. Декомпозиция может осуществляться с различными параметрами.

Тогда функция работы ИНС на примере многослойного персептрона:

$$\begin{cases} y_i^{(k)} = f_i^{(k)}\left(\sum\limits_{j=0}^{H_{k-1}} w_{ij}^{(k)} \cdot y_i^{(k-1)}\right), \\ y_i^{(0)} = x_i^{(0)} \end{cases} \tag{9}$$

где $x_i^{(0)}$ – входные данные НС, $y_i^{(k)}$ – выход i-го нейрона k-го слоя, H_{k-1} – число нейронов на слое $k-1$, $f_i^{(k)}$ – функция активации i-го нейрона k-го слоя.

Ответ: общая рекуррентная формула описания каждой из полученных блочных нейронных сетей, где массив $\left\{L_0,...,L_{D-1}\right\}$ – количество слоев нейронной сети, которые должны быть переданы на устройство с номером N:

$$\bar{X}_N : \begin{cases} \begin{cases} y_i^{(K)} = f_i^{(K)}\left(\sum\limits_{j=0}^{H_{K-1}} w_{ij}^{(K)} \cdot y_i^{(K-1)}\right), N = D-1 \\ y_i^{\left(\sum_{k=0}^{N} L_k\right)} = f_i^{\left(\sum_{k=0}^{N} L_k\right)}\left(\sum\limits_{j=0}^{H_{\sum_{k=0}^{N} L_k - 1}} w_{ij}^{\left(\sum_{k=0}^{N} L_k\right)} \cdot y_i^{\left(\sum_{k=0}^{N} L_k - 1\right)}\right), N < D-1 \end{cases} \\ \cdots \\ \begin{cases} y_i^{(L_{N-1}+1)} = y_i^{(L_{N-1})}, N > 0 \\ y_i^{(0)} = x_i^{(0)}, N = 0 \end{cases} \end{cases} \tag{10}$$

В рамках подтверждения достоверности предлагаемой математической модели потребовалось сформулировать и доказать теоремы.

Теорема 1. О существовании блочной нейронной сети. Для любого $D \le K$ существует последовательность нейронных сетей $\left\{\overline{X}_0,...,\overline{X}_{D-1}\right\}$, называемых блочными нейронными сетями, такая, что результат последователь-

ного их вычисления, т.е. когда выходное значение $y_i^{(n)}$ нейронной сети $\overline{X}_n$ является входным значением $x_i^{(n+1)}$ нейронной сети $\overline{X}_{n+1}$ для $n \in [1, D-1]$, совпадает с результатом работы исходной сети X для всех $x_i^{(0)} \in \mathbb{R}$.

Процесс преобразования каскада блочных НС в монолитную НС, результат вычисления которой совпадает с результатами блочной НС, принято называть **композицией** искусственной нейронной сети.

Теорема 2. Об эквивалентности монолитной и блочной нейронных сетей (МНС и БНС соответственно). Результаты работы монолитной нейронной сети X и последовательности блочных нейронных сетей $\{\overline{X}_0, ..., \overline{X}_{D-1}\}$, полученной декомпозицией монолитной нейронной сети X, равны для всех входных значений $x_i^{(0)} \in \mathbb{R}$.

Доказательства этих теорем полностью приведены в тексте диссертационного исследования. Адекватность математической модели подтверждается доказательствами Теорем 1 и 2.

В **третьей главе** сформулирован метод синтеза устройств реализации искусственных нейронных сетей, представлены различные способы декомпозиции нейронной сети в зависимости от входных параметров, и рассмотрена модификация метода синтеза устройств реализации искусственных нейронных сетей, позволяющая обеспечить отказоустойчивость.

Рисунок 1 – Сущность предлагаемого метода синтеза нейросетевых устройств для реализации распределенных нейронных сетей

Входными данными для реализации метода являются МНС, число устройств D, их вычислительные мощности $\overline{p}$ и флаг минимизации передаваемых данных. Метод формирует оптимальную по выбранным параметрам последовательность нейросетевых устройств для реализации режима fog computing размерности D. Сущность предлагаемого усовершенствованного метода синтеза нейросетевых устройств (Рисунок 1) заключается в том, чтобы последовательно осуществить следующие шаги:

1. Выбрать МНС и входные параметры для работы метода.
2. Осуществить декомпозиции МНС различными способами.
3. Использовать алгоритм выбора оптимального варианта декомпозиции нейронной сети для Парето-оптимизации полученных вариантов по заданным параметрам.
4. Осуществить синтез на основе оптимальной декомпозиции.

Данные между вычислительными устройствами передаются в следующем формате: байт начала, номер пакета, передаваемая информация, байт окончания. Тайм-аут ожидания следующего пакета составляет 250 мс.

Предлагается следующая классификация вариантов декомпозиции монолитной нейронной сети в зависимости от входных параметров:

1. Разделение на блоки с равным количеством слоев на узле.
2. Разделение на блоки с равным количеством нейронов на узле.
3. Разделение пропорционально производительности устройств (слои).
4. Разделение пропорционально производительности устройств (нейроны).
5. Разделение с условием минимизации передаваемых по сети данных.

В предлагаемом отказоустойчивом варианте метода устройство с номером N+1 является замещающим для устройства с номером N при условии, что замещению подвергается одно отказавшее устройство. Если требуется адаптировать каскад к потенциальному отказу двух устройств, идущих друг за другом, то устройство с номером N+2 станет замещающим для устройств N и N+1. Последнее устройство в каскаде, с номером D, не имеет устройства с номером больше своего, поэтому его диагностику и замещение берет на себя устройство с номером D-1, для этого у последнего устройства организуется обратная связь.

В **четвертой главе** описана разработка и реализация алгоритма преобразования классической нейронной сети в нейронную сеть, адаптированную для туманных вычислений, алгоритма запуска и работы распределенной нейронной сети, а также проведена оценка сложности алгоритмов. Помимо этого, разработан алгоритм выбора оптимального варианта декомпозиции нейронной сети для реализации на распределенных вычислительных устройствах.

Ограничения алгоритма декомпозиции аналогичны ограничениям математической модели. Алгоритм декомпозиции монолитной нейронной сети позволяет учитывать параметры каскада вычислительных устройств при осуществлении декомпозиции. Подробно алгоритм представлен в тексте диссертационного исследования.

Также был реализован алгоритм вычисления результатов работы нейронной сети в предложенном формате:

1. Устройство, реализующее блок с номером i, получает вектор исходных данных.
2. Обрабатывает полученный вектор с помощью блока с номером i.
3. Передает результаты обработки блоку i+1 по каналам связи.

В случае отказоустойчивой реализации предлагаемого метода результирующий вектор первого блока по дополнительно организованной связи передается на третий блок и станет входным вектором для третьего блока в случае отказа или сбоя второго блока. То есть алгоритм функционирования распреде-

ленной блочной нейронной сети в случае модифицированной отказоустойчивой реализации алгоритма с глубиной адаптации G будет последовательно реализовывать блоки второго и третьего устройств на устройстве номер три.

Разработан алгоритм выбора оптимального варианта декомпозиции нейронной сети для реализации на распределенных вычислительных устройствах (Рисунок 2).

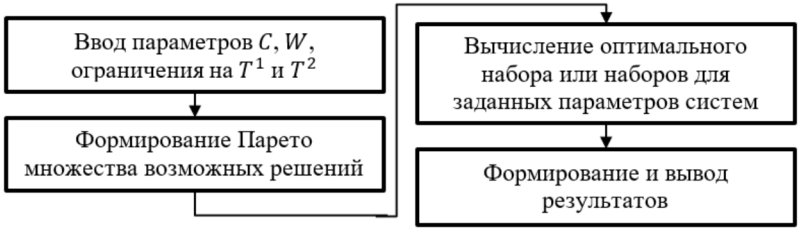

Рисунок 2 – Схема предлагаемого алгоритма выбора оптимального варианта декомпозиции нейронной сети

В процессе диссертационного исследования с целью оценки эффективности разработанного метода проведена оценка временной сложности разработанных алгоритмов: все представленные алгоритмы имеют полиномиальную сложность. Разработанные алгоритмы были реализованы в программных продуктах NNSplitter и NNImplementer на языке программирования JAVA.

В **пятой главе** выполнено схемотехническое моделирование и прототипирование, проведен анализ полученных параметров устройств, приведен пример решения четырехкритериальной задачи оптимизации.

Было осуществлено схемотехническое моделирование различных конфигураций каскадов нейросетевых устройств в САПР Proteus: монолитной НС на одном микроконтроллере и блочной НС на нескольких последовательно соединенных микроконтроллерах. Устройством, которое будет реализовывать нейросетевые вычисления, был выбрал микроконтроллер ATmega32 (Рисунок 3).

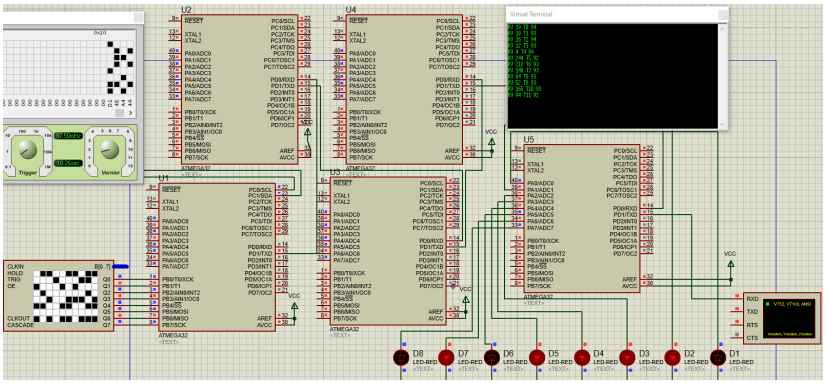

Рисунок 3 – Схема электрическая функциональная работы БНС на пяти микроконтроллерах ATMega32 в системе схемотехнического моделирования Proteus

В рамках моделирования и проведения экспериментов по запуску блочных НС осуществлялась декомпозиция 20 различных монолитных нейронных сетей. Каждая из них была декомпозирована на каскады нескольких вычислительных устройств. На вход блочной нейронной сети в каждом из моделирований направлялись 50 векторов входных параметров. Полученные результаты подтверждают, что предложенный метод синтеза нейросетевых устройств для реализации режима туманных вычислений работает верно, так как результаты во всех перечисленных случаях получаются одинаковыми. Метод синтеза позволяет получить каскад блоков блочной НС, который дает те же самые результаты, что и исходная монолитная НС, результаты моделирования показывают, что каскад из блочных нейронных сетей сохраняет основные параметры монолитной нейронной сети, из которой он был декомпозирован, например точность (Ac, Accuracy).

Для тестирования модификации метода синтеза было проведено несколько запусков различных схемотехнических моделей с дополнительными связями между устройствами и искусственно созданными отказами определенных устройств. Осуществлена декомпозиция нейронной сети на три и пять блоков. Полученные результаты подтверждают, что предложенная модификация метода синтеза нейросетевых устройств для реализации режима туманных вычислений позволяет осуществить адаптацию с глубиной G, ведь независимо от того, выполняется нейронная сеть на полностью исправном каскаде устройств или вычисления происходят в каскаде с отказом одного из вычислительных узлов, результаты в обоих случаях получаются одинаковыми.

Оценка проводилась по декомпозиции конкретной нейронной сети пятью способами. Для четырех различных конфигураций оборудования: одного без изменений, и трех с установкой дополнительного вычислительного устройства. Рассматриваемая в примере поиска оптимального решения нейронная сеть для биометрической идентификации состояла из 50 слоев и более чем 1000 математических нейронов (FingerNet, модификация модели ResNet50). Точность (Accuracy) рассматриваемой монолитной нейронной сети составляла 95,7 % [104], точность всех полученных в результате различных вариантов декомпозиции блочных НС совпала с точностью исходной нейросети.

В текущей конфигурации вычислительная система состоит из следующих элементов: микроконтроллер Atmel AT91SAM7X256-AU, коммутатор D-Link DGS-1100-05PDV2, одноплатный микрокомпьютер Raspberry PI Zero, 3-портовый управляемый коммутатор 10/100 Ethernet KSZ8993M. В одном случае предлагается добавить микроконтроллер ATMega32, во втором – еще один одноплатный компьютер Raspberry PI Zero, в третьем – Raspberry PI Model 4 B. В случае исходной архитектуры будет происходить каскадирование в режиме остаточного ресурса, то есть в рамках реализации нейросетевых вычислений будут использоваться оставшиеся относительно свободными вычислительные мощности на уже имеющихся в исходной схеме устройствах.

В других рассмотренных случаях предлагалось добавить в каскад дополнительное вычислительное устройство: в первом случае – микроконтроллер ATMega32, во втором – еще один одноплатный компьютер Raspberry PI Zero, в

третьем – Raspberry PI Model 4 B. Схема альтернативной архитектуры каскада представлена на Рисунке 4.

В этом случае были рассмотрены как варианты реализации монолитной нейронной сети одиночным добавленным вычислителем, так и варианты добавления нового вычислителя в каскад с целью распределения на него части нейросетевой нагрузки параллельно с тем, чтобы использовать ресурсы уже имеющихся устройств.

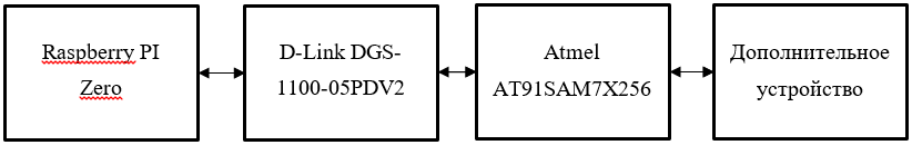

Рисунок 4 – Схема каскада ВУ, использованного во время внедрения

Каждый из четырех предложенных вариантов реализации по результатам анализа исходных данных и прототипирования вычислительных каскадов был оценен по основным параметрам оптимизации для рассматриваемой задачи: стоимость $\boldsymbol{C}$, энергопотребление $\boldsymbol{W}$, время выполнения нейросетевых вычислений на конкретном каскаде вычислителей $\boldsymbol{T^1}$ и время отклика используемого пульта голосования на поступившую извне команду $\boldsymbol{T^2}$.

Было проведено несколько запусков различных схемотехнических моделей и физических стендов. Осуществлена декомпозиция нейронной сети на блоки для исполнения на каскаде устройств в рамках каждой из предложенных схем различными способами. В случаях равномерного распределения слоев и нейронов по всем вычислительным устройствам, каждое получило примерно по 33 и 25 % вычислительной нагрузки для каскадов из трех и четырех вычислителей соответственно. В других случаях, когда нагрузка на устройства распределялась пропорционально их мощности, распределение нагрузки получилось более сложным, оно представлено в Таблице 1.

Таблица 1 – Распределение нагрузки по узлам каскада устройств

Pi Zero		D-Link		Atmel		Новое у-во		
%	Сл-в	%	Сл-в	%	Сл-в	%	Сл-в	Имя у-ва
55	27	22,5	11	22,5	11	-	-	-
44,8	22	18,4	9	18,4	9	18,4	10	ATMega32
35,5	17	14,5	7	14,5	7	35,5	19	Pi Zero
22,7	11	9,3	4	9,3	4	58,7	31	Pi 4

На вход каждой модели были последовательно представлены одинаковые исходные данные, созданные в генераторе слов. Были произведены замеры показателей времени выполнения нейросетевых вычислений и времени отклика пультов голосования. Исходя из представленных данных, можно приступить к Парето-оптимизации и выбрать наиболее оптимальный вариант для его последующего использования. На Рисунке 5 представлена точечная диаграмма, каждый из узлов – один из экспериментальных случаев по времени работы.

Таким образом, самым оптимальным вариантом в двухкритериальной оптимизации (только по времени $\boldsymbol{T^1}$ и $\boldsymbol{T^2}$) будет вариант на архитектурной схеме номер 3, с декомпозицией по 4-му способу.

Нейронная сеть была декомпозирована по 4-му способу, то есть с пропорциональным разделением нагрузки между вычислителями. Для решения четырехкритериальной задачи была построена таблица с рассматриваемыми вариантами декомпозиции и построения архитектуры вычислительного каскада (Таблица 2).

Таблица 2 – Варианты решения четырехкритериальной задачи оптимизации

Схема	C (руб.)	W (Вт)	T^1 (мс)	T^2 (мс)
Схема 1 (Исходная)	10601	6,676	86	105
Схема 2 (Добавлено у-во 1)	12211	6,811	107	96
Схема 3 (Добавлено у-во 2)	15232	7,526	85	89
Схема 4 (Добавлено у-во 3)	21501	10,176	53	80

Парето-оптимальным решением было признано решение, сохраняющее исходную архитектуру вычислительного каскада (Схема 1, черный сектор на Рисунке 5), поскольку оно позволяет избежать значительных материальных затрат на закупку оборудования (параметр стоимости C) и увеличения суммарной мощности вычислительного каскада (параметр W), сохраняя при этом допустимые скорости отклика устройств и время выполнения нейросетевых вычислений (параметры T^1 и T^2).

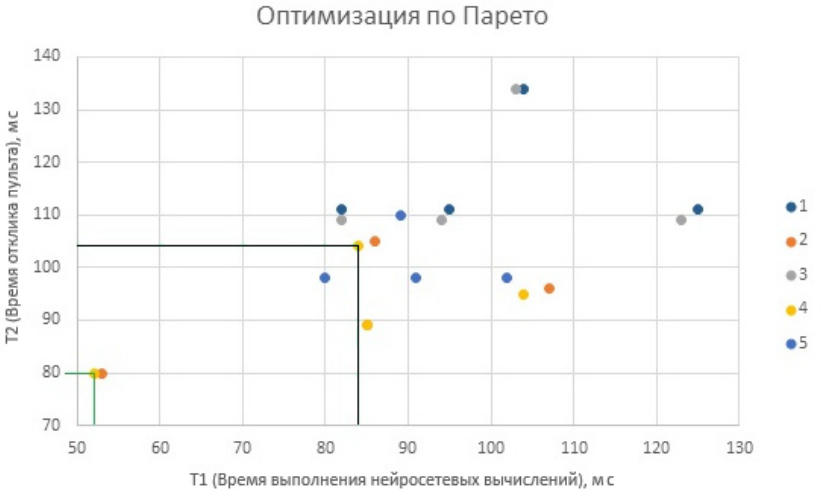

Рисунок 5 – Парето-оптимальные решения в двухкритериальной задаче (малый сектор) и в четырехкритериальной задаче (большой сектор)

На Рисунке 5 представлены декомпозиции монолитной нейронной сети пятью способами для всех схем каскадов физических устройств. Оптимальным

вариантом в предлагаемой модели оптимизации по требуемым критериям является вариант с сохранением исходной архитектуры и декомпозицией нейронной сети на блоки способом номер 4, то есть с разделением на блоки пропорционально мощности с точностью до нейронов на слоях. Результаты говорят о том, что в рамках предложенного метода и различных способов его реализации возможно провести оптимизацию и выбрать оптимальную декомпозицию НС, что подтверждает эффективность предложенного метода.

Внедрение полученного метода синтеза нейросетевых устройств происходило на базе ЗАО «Проминформ», разрабатывающего системы голосования для высших органов государственной власти (Государственная Дума, Совет Федерации). Метод был использован для прототипирования системы распределения вычислительных мощностей каскада устройств для выполнения вычислений нейронной сети, осуществляющей обработку дактилоскопических данных, полученных от специализированного сенсора с целью биометрической идентификации депутата на рабочем месте.

В приложениях представлены: листинг алгоритмов декомпозиции монолитной нейронной сети и запуска блочной нейронной сети, акты внедрения научных результатов в ЗАО Проминформ и ФГБОУ ВО ПНИПУ, свидетельства о государственной регистрации программ для ЭВМ.

ЗАКЛЮЧЕНИЕ

Представленная диссертационная работа посвящена решению важной научно-технической проблемы – улучшению эксплуатационно-технических показателей вычислительных систем и их элементов на основе декомпозиции искусственной нейронной сети и реализации полученных блоков в каскаде нейросетевых устройств. В диссертационной работе поставлены и решены следующие задачи исследования:

1. Разработана *математическая модель искусственной нейронной сети для синтеза нейросетевых устройств, ориентированных на туманные вычисления*. Это *позволило* учесть требуемую загрузку вычислительных узлов при распределении блоков нейронной сети между различными устройствами.

2. Разработан *метод синтеза устройств реализации искусственных нейронных сетей, ориентированных на туманные вычисления, и его модификация, обеспечивающая отказоустойчивость*. Это *позволило* значительно снизить нагрузку на те вычислительные узлы, которые требовалось разгрузить в рамках рассматриваемой задачи, и организовать вычислительный каскад таким образом, чтобы продолжать работу даже в случае отказа части устройств в каскаде.

3. Разработан *алгоритм декомпозиции монолитной нейронной сети на каскад блоков блочной нейронной сети*, *адаптированной для туманных вычислений*. Это *позволит* проводить многократную декомпозицию в глубину, например, если потребуется декомпозировать отдельный блок еще на несколько блоков.

4. Разработан *алгоритм выбора оптимального варианта декомпозиции нейронной сети для реализации на распределенных вычислительных устройствах*.

Это *позволило* найти оптимальную декомпозицию монолитной нейронной сети сразу по нескольким важным для вычислительной системы параметрам, за счет чего стало возможным расширение обработки данных вычислительной системой без увеличения ее стоимости.

Дальнейшие исследования целесообразны в области декомпозиции нейронных сетей с другими актуальными нейросетевыми архитектурами.

Практическая значимость диссертационного исследования заключается в экономии вычислительных ресурсов определенных узлов за счет распределения нагрузки на менее загруженные части системы, а также повышения отказоустойчивости системы за счет осуществления диагностики и реконфигурации вычислительных узлов и уменьшении затрат материальных ресурсов для реализации распределенных каскадов нейросетевых устройств, что подтверждается актом внедрения. Разработанные программные продукты прошли тестирование и получили свидетельства о государственной регистрации.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИОННОЙ РАБОТЫ

Публикации в ведущих рецензируемых научных изданиях

1. **Бахтин, В.В.** Модификация алгоритма идентификации и категоризации научных терминов с использованием нейронной сети / В.В. Бахтин // Нейрокомпьютеры: разработка, применение. – 2019. – Т. 21, № 3. – С. 14–19.

2. **Бахтин, В.В.** Математическая модель искусственной нейронной сети для устройств на ПЛИС и микроконтроллерах, ориентированных на туманные вычисления / В.В. Бахтин // Вестник Пермского национального исследовательского политехнического университета. Электротехника. Информационные технологии, системы управления. – 2021. – № 40. – С. 109–129.

3. **Бахтин, В.В.** Алгоритм разделения монолитной нейронной сети для реализации туманных вычислений в устройствах на программируемой логике / В.В. Бахтин // Вестник Пермского национального исследовательского политехнического университета. Электротехника. Информационные технологии, системы управления. – 2022. – № 41. – С. 123–145.

4. **Бахтин, В.В.** Метод синтеза устройств нейросетевого распознавания на программируемой логике для реализации режима fog computing / В.В. Бахтин, С.Ф. Тюрин, И.А. Подлесных // Вестник Пермского национального исследовательского политехнического университета. Электротехника. Информационные технологии, системы управления. – 2022. – № 41. – С. 168–188.

5. **Бахтин, В.В.** Решение задачи многокритериальной оптимизации вариантов декомпозиции нейронной сети и компоновки каскада вычислительных устройств методом Парето / В.В. Бахтин, И.А. Подлесных, С.Ф. Тюрин // Вестник Пермского национального исследовательского политехнического университета. Электротехника. Информационные технологии, системы управления. – 2022. – № 43. – С. 136–156.

Публикации в изданиях, индексируемых в международной базе цитирования Scopus

6. **Bakhtin, V.V.** Algorithm for Decomposition of a Monolithic Neural Network into a Cascade of Block Neural Networks for the Fog Computing / V.V. Bakhtin // 2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). – 2022. – P. 238–241. DOI: 10.1109/ElConRus54750.2022.9755533

7. Podlesnykh, I.A. Mathematical Model of a Recurrent Neural Network for Programmable Devices Focused on Fog Computing / I.A. Podlesnykh, **V.V. Bakhtin** // 2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). – 2022. – P. 395–397. DOI: 10.1109/ElConRus54750.2022.9755677

8. **Bakhtin, V.V.** New TSBuilder: Shifting towards Cognition / V.V. Bakhtin, E.V. Isaeva // 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). – 2019. – P. 179–181. DOI: 10.1109/EIConRus.2019.8656917

Свидетельства о государственной регистрации программ для ЭВМ

9. Программа для ЭВМ № 2022611627 Российская Федерация. Программный продукт «NNSplitter» для декомпозиции монолитных НС на каскад блочных НС для синтеза нейросетевых устройств на программируемой логике : № 2022611627 : заявл. 19.01.2022 : опубл. 28.01.2022 / **Бахтин В.В.** – 1 с. – Текст : непосредственный.

10. Программа для ЭВМ № 2022615562 Российская Федерация. Программный продукт «NNImplementer» для реализации запуска и работы блочной НС на каскаде нейросетевых устройств на программируемой логике : № 2022615562 : заявл. 27.03.2022 : опубл. 31.03.2022 / **Бахтин В.В.** – 1 с. – Текст : непосредственный.

Публикации в прочих изданиях, в том числе материалы конференций

11. **Бахтин, В.В.** Исследование декомпозиции нейронной сети в системе схемотехнического моделирования Proteus / В.В. Бахтин, И.А. Подлесных, С.Ф. Тюрин // Вестник Пермского университета. Математика. Механика. Информатика. – 2022. – № 2 (57). – С. 73–80.

12. **Бахтин, В.В.** Алгоритм построения графа совместной работы каскадов устройств нейросетевого распознавания, реализующих блочные нейронные сети / В.В. Бахтин, И.А. Подлесных // Сборник материалов IX Международной научной конференции, посвященной 85-летию профессора В.И. Потапова. – Омск, 2021. – С. 277–278.

13. Подлесных, И.А. Методы декомпозиции искусственных нейронных сетей с учетом возможности распараллеливания вычислений / И.А. Подлесных, **В.В. Бахтин** // Автоматизированные системы управления и информационные тех-

нологии: сборник материалов всероссийской научно-технической конференции. – Пермь, 2022. – Т. 1. – С. 215–220.

14. Подлесных, И.А. Усовершенствование метода проектирования нейросетевых устройств для туманных вычислений / И.А. Подлесных, **В.В. Бахтин** // Инновационные технологии: теория, инструменты, практика: тезисы XIV Международной интернет-конференции молодых ученых, аспирантов и студентов (InnoTech-2022).

15. **Bakhtin, V.V.** TSBuilder 2.0: Improving the Identification Accuracy Due to Synonymy / V.V. Bakhtin, E.V. Isaeva, A.V. Tararkov // 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). – 2020. – P. 225–228. DOI: 10.1109/EIConRus49466.2020.9039207

16. **Bakhtin, V.V.** TSMiner: from TSBuilder to Ecosystem / V.V. Bakhtin, E.V. Isaeva, A.V. Tararkov // 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). – 2021. – P. 221–224. DOI: 10.1109/ElConRus51938.2021.9396569

17. **Bakhtin, V.** Developing an Algorithm for Identification and Categorization of Scientific Terms in Natural Language Text through the Elements of Artificial Intelligence / V. Bakhtin, E. Isaeva // 14th International Scientific-Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE) – 44894. Proceedings. – Novosibirsk, 2018. – P. 384–390.

18. Isaeva, E. Collecting the Database for the Neural Network Deep Learning Implementation / E. Isaeva, **V. Bakhtin**, A. Tararkov // Digital Science. DSIC18 2018. Advances in Intelligent Systems and Computing. – 2019. – Vol. 850. – P. 12–18. – Springer, Cham, 2019. DOI: 10.1007/978-3-030-02351-5_2

19. Isaeva, E. Formal Cross-Domain Ontologization of Human Knowledge / E. Isaeva, **V. Bakhtin**, A. Tararkov // Information Technology and Systems. ICITS 2020. Advances in Intelligent Systems and Computing. – 2020. – Vol. 1137. – P. 94–103. – Springer, Cham, 2020. DOI: 10.1007/978-3-030-40690-5_10

20. Isaeva, E. Ontologization and Term System Modelling by means of AI Methods/ E. Isaeva, A. Tararkov, **V. Bakhtin** // Specialized Knowledge Mediation. – 2022. – P. 139–149. – Springer, Cham, 2022. DOI: 10.1007/978-3-030-95104-7_7

Подписано в печать 27.04.2023. Формат 60×90/16.
Усл. печ. л. 1,3. Тираж 100 экз. Заказ № 092/2023.

Отпечатано в типографии
Издательства Пермского национального
исследовательского политехнического университета.
Адрес: 614990, г. Пермь, Комсомольский проспект, 29, к. 113.
Тел. (342) 219-80-33.