THE MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION



ISSN 2687-0517

Computing, Telecommunications and Control

Vol. 19, No. 2
2026

Peter the Great St. Petersburg
Polytechnic University
2026

EDITORIAL COUNCIL

Prof. Dr. *Dmitry G. Arseniev* corresponding member of RAS, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Vladimir S. Zaborovsky*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Dmitry P. Zegzhda*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Vladimir N. Kozlov*, Peter the Great St. Petersburg Polytechnic University, Russia;
Assoc. Prof. Dr. *Ivan S. Mukhin*, Alferov University, St. Petersburg, Russia;
Prof. Dr. *Igor G. Chernorutsky*, Peter the Great St. Petersburg Polytechnic University, Russia.

EDITORIAL BOARD

Editor-in-chief
Prof. Dr. *Alexander S. Korotkov*, Peter the Great St. Petersburg Polytechnic University, Russia;

Members:
Assoc. Prof. Dr. *Pavel D. Drobintsev*, Peter the Great St. Petersburg Polytechnic University, Russia;
Assoc. Prof. Dr. *Vladimir M. Itsykson*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Philippe Ferrari*, Grenoble Alpes University, France;
Prof. Dr. *Yevgeni Koucheryavy*, Tampere University of Technology, Finland;
Prof. Dr. *Wolfgang Krautschneider*, Hamburg University of Technology, Germany;
Prof. Dr. *Fa-Long Luo*, University of Washington, USA;
Prof. Dr. *Sergey B. Makarov*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Emil Novakov*, Grenoble Alpes University, France;
Prof. Dr. *Nikolay N. Prokopenko*, Don State Technical University, Russia;
Prof. Dr. *Mikhail G. Putrya*, National Research University of Electronic Technology, Russia;
Sen. Assoc. Prof. Dr. *Evgeny Pyshkin*, University of Aizu, Japan;
Prof. Dr. *Viacheslav P. Shkodyrev*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Vladimir A. Sorotsky*, Peter the Great St. Petersburg Polytechnic University, Russia
Prof. Dr. *Igor A. Tsikin*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Sergey M. Ustinov*, Peter the Great St. Petersburg Polytechnic University, Russia;
Prof. Dr. *Lev V. Utkin*, Peter the Great St. Petersburg Polytechnic University, Russia.

The journal is included in the List of Leading PeerReviewed Scientific Journals and other editions to publish major findings of PhD theses for the research degrees of Doctor of Sciences and Candidate of Sciences.

Open access journal is to publish articles of a high scientific level covering advanced experience, research results, theoretical and practical problems of informatics, electronics, telecommunications, and control.

The journal is indexed by Ulrich's Periodicals Directory, Google Scholar, EBSCO, ProQuest, Index Copernicus, VINITI RAS Abstract Journal (Referativnyi Zhurnal), VINITI RAS Scientific and Technical Literature Collection, Russian Science Citation Index (RSCI) database Scientific Electronic Library and Math-Net.ru databases.

The journal is registered with the Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications (ROSKOMNADZOR). Certificate ЭЛ No. ФС77-77378 issued 25.12.2019.

Editorial office

Dr. Sc., Professor A.S. Korotkov – Editor-in-Chief;

Ph.Ch.S. Bastian – literary editor, proofreader; G.A. Pyshkina – editorial manager; A.A. Kononova – computer layout; I.E. Lebedeva – English translation.

Address: 195251 Polytekhnicheskaya Str. 29, St. Petersburg, Russia.

+7 (812) 552-6216, e-mail: infocom@spbstu.ru

Release date: 30.06.2026

© Peter the Great St. Petersburg Polytechnic University, 2026

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ



ISSN 2687-0517

Информатика, телекоммуникации и управление

Том 19, № 2
2026

Санкт-Петербургский политехнический
университет Петра Великого
2026

ИНФОРМАТИКА, ТЕЛЕКОММУНИКАЦИИ И УПРАВЛЕНИЕ

РЕДАКЦИОННЫЙ СОВЕТ ЖУРНАЛА

Арсеньев Д.Г., чл.-кор. РАН, д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия; *Заборовский В.С.*, д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия; *Зегжда Д.П.*, чл.-кор. РАН, д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия; *Козлов В.Н.*, д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия; *Мухин И.С.*, д-р физ.-мат. наук, доцент, Санкт-Петербургский национальный исследовательский Академический университет им. Ж.И. Алферова Российской академии наук, Санкт-Петербург, Россия; *Черноруцкий И.Г.*, д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия.

РЕДАКЦИОННАЯ КОЛЛЕГИЯ ЖУРНАЛА

Главный редактор

Коротков А.С., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Редакционная коллегия:

Дробинцев П.Д., канд. техн. наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Ицыксон В.М., канд. техн. наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Феррари Ф., профессор, Университет Гренобль-Альпы, Гренобль, Франция;

Краутшнайдер В., профессор, Гамбургский технический университет, Гамбург, Германия;

Кучерявый Е.А., канд. техн. наук, профессор, Университет Тампере, Финляндия.

Люо Ф.-Л., University of Washington, Washington, USA;

Макаров С.Б., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Новаков Э., профессор, Университет Гренобль-Альпы, Гренобль, Франция;

Прокопенко Н.Н., д-р техн. наук, профессор, Донской государственный технический университет, г. Ростов-на-Дону, Россия;

Путря М.Г., д-р техн. наук, профессор, Национальный исследовательский университет «Московский институт электронной техники», Москва, Россия;

Пышкин Е.В., профессор, Университет Айзу, Айзу-Вакаматсу, Япония;

Сороцкий В.А., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Устинов С.М., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Уткин Л.В., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Цикин И.А., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия;

Шкодырев В.П., д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия.

Журнал с 2002 года входит в Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени доктора и кандидата наук.

Сетевое издание открытого доступа публикует статьи высокого научного уровня, освещающие передовой опыт, результаты НИР, теоретические и практические проблемы информатики, электроники, телекоммуникаций, управления.

Сведения о публикациях представлены в Реферативном журнале ВИНИТИ РАН, в международной справочной системе «Ulrich`s Periodical Directory», в Российской государственной библиотеке. В базах данных: Российский индекс научного цитирования (РИНЦ), Google Scholar, EBSCO, Math-Net.Ru, ProQuest, Index Copernicus.

Журнал зарегистрирован Федеральной службой по надзору в сфере информационных технологий и массовых коммуникаций (Роскомнадзор). Свидетельство о регистрации ЭЛ № ФС77-77378 от 25.12.2019.

Учредитель и издатель: Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация.

Редакция журнала

д-р техн. наук, профессор А.С. Коротков – главный редактор;

Ф.К.С. Бастиан – литературный редактор, корректор; Г.А. Пышкина – ответственный секретарь, выпускающий редактор;

А.А. Кононова – компьютерная вёрстка; И.Е. Лебедева – перевод на английский язык.

Адрес редакции: Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29.

Тел. редакции +7(812) 552-62-16, e-mail редакции: infocom@spbstu.ru

Дата выхода: 30.06.2026

© Санкт-Петербургский политехнический университет Петра Великого, 2026

Contents

Intelligent Systems and Technologies, Artificial Intelligence

Skrebenkov D.I., Budanov D.O. Spiking neural network analog two-variable neuron implementation ... 7

Guo C., Potekhin V.V. TFCA-GRU: A temporal and feature cooperative attention GRU for remaining useful life prediction of bearings across conditions 16

Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Verbova N.M. Survival analysis using Mahalanobis distance in Kernels of the Beran estimator 31

Circuits and Systems for Receiving, Transmitting and Signal Processing

Pilipko M.M., Morozov D.V. A frequency-selective circuit in 180 nm CMOS 44

Klimenko D.V., Nikitin A.B., Stroganov A.A. Design of power limiter for antenna-feeder system for direction finding protection on GaAs pHEMT 52

Pilipko M.M., Morozov D.V. A spiking neural network in 180 nm CMOS 61

Simulations of Computer, Telecommunications and Control Systems

Chernyi V.G., Bolsunovskaya M.V. Analysis of the influence of weighting coefficients in a significance metric for 3D scene objects in preliminary rendering optimization setup 70

System Analysis and Control

Shaidullin R.A. Method for implementing scheduling systems in discrete-continuous production: the case of oil refineries 81

Deeb D.A., Khokhlovskiy V.N. Structure-aware decision support system for MOO-driven control 95

Fershtadt M.I., Shashikhin V.N. Adaptive weight adjustment in multicriteria control of nonlinear systems based on a Lyapunov function 109

Содержание

Интеллектуальные системы и технологии, искусственный интеллект

Скребенков Д.И., Буданов Д.О. Реализация аналогового двоичного нейрона для импульсной нейронной сети 7

Го Ч., Потехин В.В. TFCA-GRU: алгоритм GRU с временным и характеристическим кооперативным вниманием для прогнозирования остаточного срока службы подшипников в различных условиях 16

Уткин Л.В., Константинов А.В., Кирпиченко С.Р., Вербова Н.М. Анализ выживаемости с использованием расстояния Махаланобиса в ядрах модели Берана 31

Устройства и системы передачи, приема и обработки сигналов

Пилипко М.М., Морозов Д.В. Частотно-избирательное устройство по технологии КМОП 180 нм 44

Клименко Д.В., Никитин А.Б., Строганов А.А. Разработка ограничителя мощности для защиты антенно-фидерных систем пеленгования на основе GaAs pHEMT-технологии 52

Пилипко М.М., Морозов Д.В. Импульсная нейронная сеть по технологии КМОП 180 нм 61

Моделирование вычислительных, телекоммуникационных и управляющих систем

Чёрный В.Г., Болсуновская М.В. Анализ влияния коэффициентов метрики значимости объектов 3D-сцены для предварительной настройки оптимизации рендеринга 70

Системный анализ и управление

Шайдуллин Р.А. Метод внедрения систем календарного планирования в дискретно-непрерывных производствах: пример НПЗ 81

Диб А., Хохловский В.Н. Структурно-ориентированная система поддержки принятия решений для управления, основанного на многокритериальной оптимизации 95

Ферштадт М.И., Шашихин В.Н. Адаптивная настройка весов в задаче многокритериального управления нелинейными системами с использованием функции Ляпунова 109

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 7–15.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 7–15.

Intelligent Systems and Technologies, Artificial Intelligence

Интеллектуальные системы и технологии, искусственный интеллект

Research article
DOI: https://doi.org/10.18721/JCSTCS.19201
UDC 004.032.26

SPIKING NEURAL NETWORK ANALOG TWO-VARIABLE NEURON IMPLEMENTATION

D.I. Skrebenkov ✉ , *D.O. Budanov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ skrebenkov_di@spbstu.ru

Abstract. Currently, the problem of using neural networks in embedded systems, which have constraints on power consumption and area occupied, is relevant. At the same time, there is growing interest in spiking neural networks – more biologically realistic neural networks with potentially higher energy efficiency. This paper presents a review of existing hardware implementations of spiking neural networks, implemented both in the digital and analog domains. A hardware implementation of analog spiking neuron and synapse fabricated in Mikron JSC 180 nm CMOS technology with a reduced supply voltage of 0.6 V is presented. According to the simulation results, the energy consumption of a single neuron was 1.46 pJ/spike. The paper also provides a formula for comparing energy consumption between neurons fabricated in different technologies and with different supply voltages. According to this formula, the energy consumption of the implemented neuron is comparable to that of the referenced sources.

Keywords: spiking neural network, power consumption, spiking neuron, analog neuron, analog synapse

Citation: Skrebenkov D.I., Budanov D.O. Spiking neural network analog two-variable neuron implementation. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 7–15. DOI: 10.18721/JCSTCS.19201

© Skrebenkov D.I., Budanov D.O., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19201
УДК 004.032.26

РЕАЛИЗАЦИЯ АНАЛОГОВОГО ДВОИЧНОГО НЕЙРОНА ДЛЯ ИМПУЛЬСНОЙ НЕЙРОННОЙ СЕТИ

Д.И. Скребенков ✉ , *Д.О. Буданов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ skrebenkov_di@spbstu.ru

Аннотация. В настоящее время актуальна проблема использования нейронных сетей во встраиваемых системах, в которых существуют ограничения по потребляемой мощности и занимаемой площади. В то же время растет интерес к импульсным нейронным сетям – более биологически реалистичным нейронным сетям с потенциально более высокой энергоэффективностью. В данной работе представлен обзор существующих аппаратных реализаций импульсных нейронных сетей, реализованных как в цифровой области, так и в аналоговой. Представлена аппаратная реализация аналоговых импульсных нейрона и синапса, выполненного по технологии АО «Микрон» 180 нм КМОП с пониженным напряжением питания 0,6 В. По результатам моделирования потребляемая энергия одного нейрона составила 1,46 пДж/импульс. Также в работе приведена формула для сравнения потребляемой энергии между нейронами, выполненными по разным технологиям и с разным напряжением питания. По данной формуле потребляемая энергия реализованного нейрона сопоставима со сравниваемыми источниками.

Ключевые слова: импульсная нейронная сеть, потребляемая мощность, импульсный нейрон, аналоговый нейрон, аналоговый синапс

Для цитирования: Skrebenkov D.I., Budanov D.O. Spiking neural network analog two-variable neuron implementation // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 7–15. DOI: 10.18721/JCSTCS.19201

Introduction

Artificial neural networks (ANNs) are widely used in various areas such as classification and clustering, image and speech recognition, prediction etc. Several types of neural network are developed, for example, feed-forward neural network (FFNN), convolutional neural network (CNN), recurrent neural network (RNN) etc. ANNs can also be applied in embedded systems for vehicle driver assistance [1], real-time image recognition [2, 3], monitoring [4] etc. The use of ANNs in embedded system is limited by small area and low power consumption that leads to the need to develop specific ANNs which can meet these restrictions. Spiking neural network (SNN) uses a different type of neuron interaction compared to other ANNs. The nature of this network can be used to develop neural networks with lower power consumption that are suitable for embedded devices.

This paper describes the basics of SNNs and shows several existing approaches to spiking neuron implementations. One of the approaches has been implemented using 180 nm technology to evaluate its benefits and drawbacks.

SNN basics

Basic principle

An SNN is a next-generation of neural network that is mostly based on biological mechanisms of neuron interactions. An SNN operates with spikes – short impulses which can activate a neuron and

© Скребенков Д.И., Буданов Д.О., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого



Fig. 1. Spiking neuron model [5]

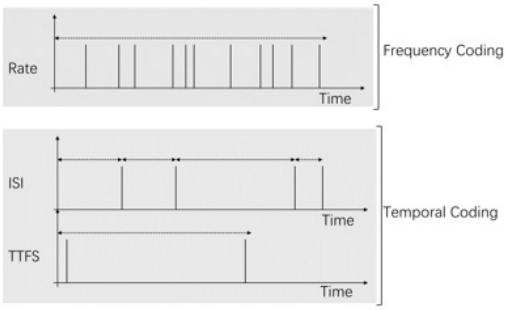


Fig. 2. Frequency and temporal encoding [5]

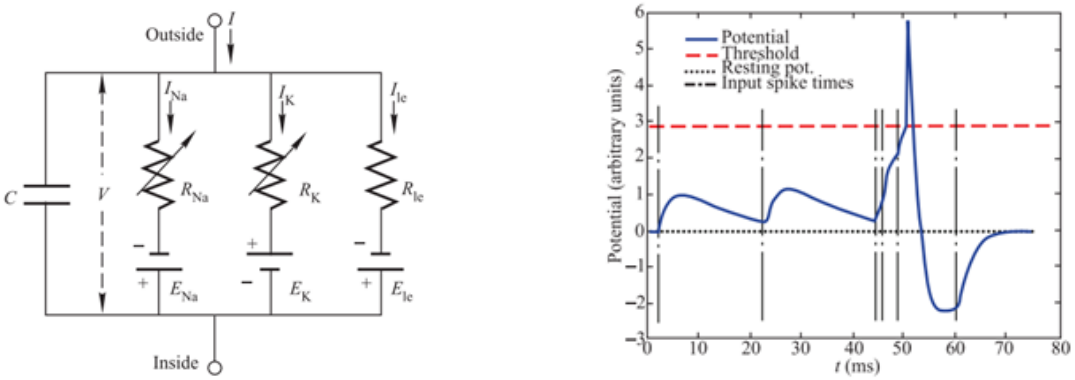


Fig. 3. Equivalent circuit and time diagram of HH model [6]

make it produce another spike. An illustrative example of a spiking neuron performance is shown in Fig. 1.

Encoding

Data encoding in SNNs is also different compared to ANNs. An SNN operates with data distributed in time, so encoding may be done by the number of spikes in a given period (frequency encoding) or by the time between spikes (temporal encoding). These types are shown in Fig. 2.

Neuron models

There are a lot of spiking neuron models. Some of them are purely mathematical models that are not applicable in hardware-implemented neural networks because of their complex computations, such as the Izhikevich model [5]. Another popular model is the Hodgkin-Huxley (HH) model [6] that includes several biological parameters of a biological neuron. The equivalent circuit and time diagram of the HH neuron model are shown in Fig. 3.

The most popular spiking neuron model is the Leaky Integrate-and-Fire (LIF) neuron [6]. In comparison with the HH model, it does not have ion potentials, but it still considers several biological mechanisms such as refractory period and leakage when there is no input excitation. All observed hardware implementations of SNN use this model of neuron due to its balance between simplicity and biological plausibility. The Integrate-and-Fire (IF) neuron model [6] does not have a refractory period and is one of the simplest spiking neuron models. However, according to a review of papers devoted to hardware implementation of neural networks, the IF model is not used as widely as the LIF one.

Hardware implementation of SNN

General device implementation

Despite focusing on hardware implementation, it is necessary to consider software-based implementations using general-purpose devices for computations such as CPUs and GPUs. In [7], a comparison between an ANN and an SNN for a classification task is presented. The data for classification were real-time ECG signals. The paper shows that the SNN accuracy was 2.5% higher than that of the ANN. However, there is no additional data in the paper to reveal which structure of ANN was used. In [8], an SNN for ECG signals processing is presented. The paper shows that the applied SNN can achieve 93–94% accuracy for real-time ECG classification.

FPGA-based implementation

FPGA is a widely used device for developing accelerators for ANN because of its high flexibility. In [9], both a new type of data encoding for SNNs and important parameters for embedded devices are presented. According to this paper, FPGA solutions in recent years can achieve 97% classification accuracy with power consumption down to 100 mW. In [10], an algorithm to switch from an ANN to an SNN to avoid the training operation for SNN is shown, which is usually more complex than training an ANN. A comparison between the starting ANN and the resulting SNN shows that the accuracy slightly decreased from 92.43 to 91.96% and power consumption decreased significantly from 0.56 to 0.20 mW.

ASIC digital implementation

When it comes to ASIC, there is a possibility to use not just a digital domain but also mixed and analog circuits to implement neural networks.

In [11], digital SNN was firstly verified using FPGA and then implemented in ASIC using 40 nm technology and a comparison of several recent papers was given. According to this comparison, the accuracy of implemented SNN achieves 90% or higher and the occupied area varies from 0.1 mm^2 to 40 mm^2. Energy consumption in the highlighted work achieve 0.75 μJ per classification. In [12], multicore coprocessor was implemented using 14 nm technology and achieves 26.7 pJ/spike energy consuming with a total occupied area 60 mm^2.

ASIC analog implementation

In [13], a full analog neuron circuit for SNN with clocked comparators is presented. This solution is implemented using 22 nm technology and achieves 87 fJ/spike energy consumption with 14000 um^2 occupied area. This work also considers a comparison between several mixed-signal approaches which shows that energy consumption can achieve values from hundreds of fJ/spike to tens of pJ/spike. In [14], an analog neuron implemented in 150 nm technology with energy consumption 55 fJ/spike and occupied area 250 um^2 is presented. It shows noticeable energy consumption but requires high-voltage additional circuits in training stage.

SNN neuron implementation

A spiking neuron and a synapse were implemented based on [15]. Their performance was then evaluated. The circuit diagram of the neuron is shown in Fig. 4.

The neuron has three units. The first unit calculates the membrane potential using a current mirror and a capacitor. The input current drives the current mirror when the input current is higher than the

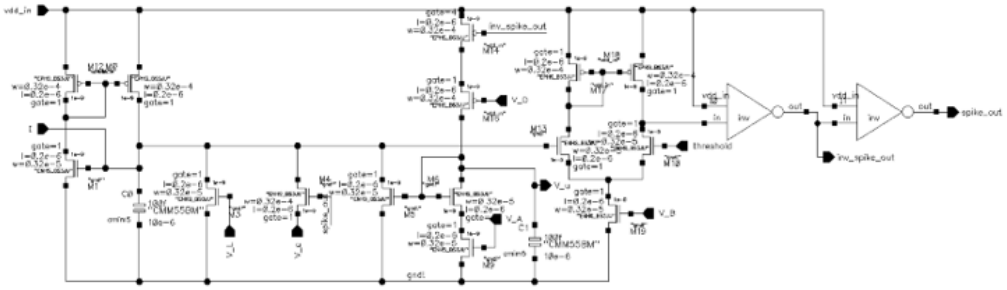


Fig. 4. Neuron circuit

leakage one. The membrane potential is proportional to the charge stored in capacitor C0 and decays through transistors M3 and M5. Transistor M5 is controlled by external voltage V_L and provides the leakage current. When a spike is generated, transistor M4 opens and the membrane potential is set to the value of external voltage V_c.

The second unit calculates the adaptation potential and its value is proportional to the charge stored in capacitor C1. When a spike is generated, transistor M14 opens and the adaptation potential is increased depending on external voltage V_D, which is the gate voltage of transistor M16. When a spike is not generated, the adaptation potential decays through transistors M6 and M9, and the latter is controlled by external voltage V_A. The current flow of transistor M5 increases as the adaptation potential increases and makes it more difficult for the membrane potential to increase.

The third unit is a comparator which is controlled by external voltages for setting the threshold and bias voltages. The circuit diagram of the synapse is shown in Fig. 5.

The neuron and the synapse were implemented using Mikron JSC 180 nm CMOS technology. The supply voltage was decreased from 1.8 V, which is standard for the technology, to 0.6 V. It allows power consumption to be reduced by moving the operating points of transistors to the sub-threshold region, as is done in the reference paper.

The simulation circuit is shown in Fig. 6.

The input source is implemented as a PMOS current mirror. The value of the input current is controlled by transistor M3 with a DC gate voltage. A pulse source is used to set the initial state. V_B voltage is set to Vdd to bring the comparator to its operating point. V_D is a PMOS gate voltage, so it is set low to let current flow when a spike is generated. V_c is a relaxation voltage, so it is also set low. V_L and V_A are set high (400 mV and 600 mV respectively) to achieve a noticeable leakage current. This simulation contains only one neuron whose output is the output of inverters, and capacitors were set as a load for it. During the simulation, the power consumption for each spike was measured, and its value is 1.46 pJ per spike.

For comparison between neurons implemented with different technologies, technology-independent energy estimation per spike is given by:

$$W = \frac{W0}{\left(\frac{T0}{T_{ref}}\right)^2 \times \left(\frac{Vdd0}{Vdd_{ref}}\right)},$$

where W is the technology-independent energy per spike, $W0$ is the technology-dependent energy per spike, $T0$ is the technology node of the evaluated circuit, T_{ref} is the reference technology node, $Vdd0$ is the voltage supply of estimated circuit and Vdd_{ref} is the reference voltage supply.

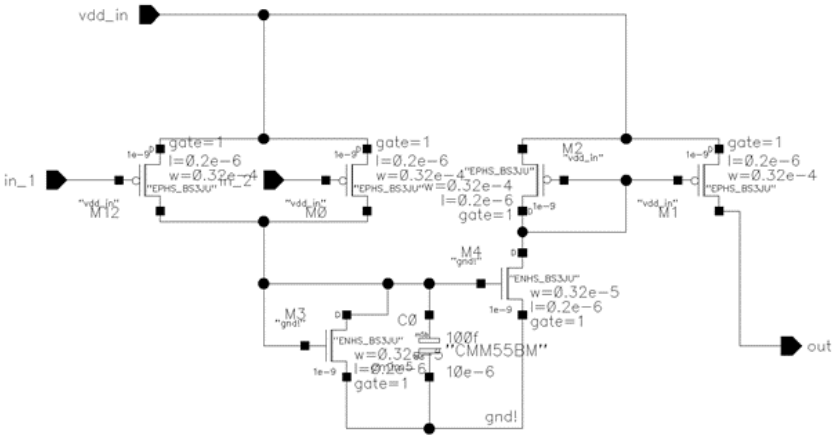


Fig. 5. Synapse circuit

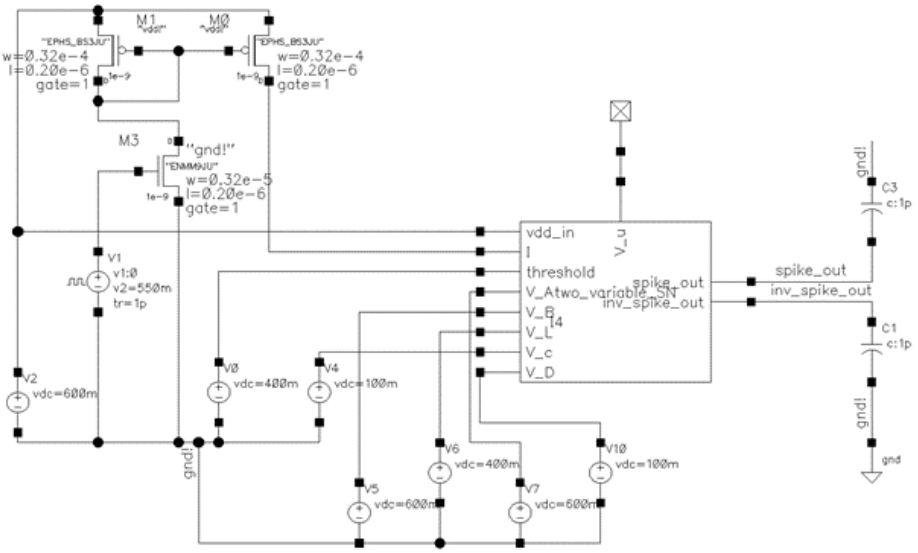


Fig. 6. Simulation circuit

According to this estimation, the neuron circuits from the references are compared with the implemented neuron circuit. Results are shown in Table 1.

Table 1

Comparison of estimated energy consumption

Neuron	Technology node	Vdd, V	Estimated W, pJ per spike
[13]	22 nm	0.8	13.06
[14]	150 nm	0.4	0.362
[15]	65 nm	0.3	4.017
This work	180 nm	0.6	4.424

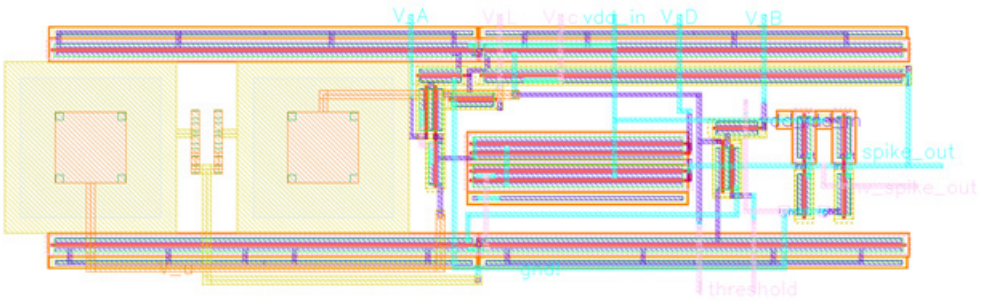


Fig. 7. SNN neuron layout

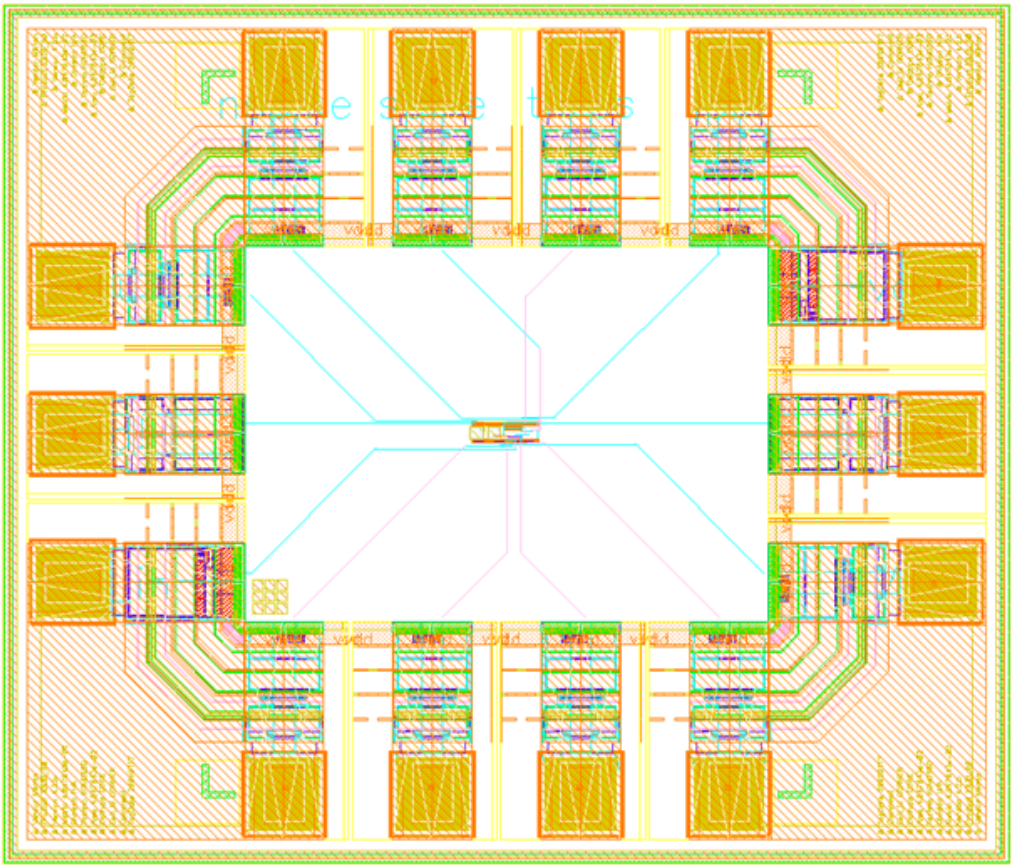


Fig. 8. Neuron with ESD guard ring

After schematic simulation, the layout of neuron circuit was designed. The layout of the SNN neuron is shown in Fig. 7. The occupied area of the neuron is 1440 μm^2. A guard ring was implemented around the chip to provide a protection against electrostatic discharge and to enable experimental measurements. The total area of the chip with the ESD guard ring is 887400 μm^2.

A synapse with two input neurons (with same weight values) and one output neuron was simulated taking into account the parasitic parameters of the SNN neuron circuit layout. A significant drawback of this synapse is that the weights of the neurons are set through the width of the transistors (M12 and M0 in Fig. 5). This will lead to weight mismatch on a real device because the effective width of the

transistors is depending on several factors of environment where this device is used and on semiconductor process effects during chip manufacturing. For further applications, it is necessary to move away from setting the weights using the transistor width and provide another method of weight setting for the SNN based on this neuron.

Conclusion

This work has presented a hardware implementation of an analog SNN neuron and a synapse with neuron energy consumption 1.46 pJ/spike at a supply voltage of 0.6 V in Mikron JSC 180 nm CMOS technology. Using the formula for comparison between implementations with different technologies, it was shown that the implementation presented in this work has comparable energy consumption compared to the reviewed works. A circuit with several neurons and their connection via a synapse was simulated to validate its performance. A drawback of the implementation presented in this work is that the neurons weights are set by the width of the synapse transistors, which will lead to weight mismatch. Further studies are required to find other ways of setting neurons weights and to reduce energy consumption of the neuron circuit for better energy efficiency.

REFERENCES

1. **Barnwal R., Srivastava R., Vimalathithan R., Kala S.** Advanced Driver Assistance System for Autonomous Vehicles Using Deep Neural Network. *2022 IEEE 10th Region 10 Humanitarian Technology Conference* (*R10-HTC*), 2022, Pp. 342–347. DOI: 10.1109/R10-HTC54060.2022.9929654

2. **Zhang M., Liao W., Zhang J., Gao H., Wang F., Lin B.** Embedded Face Recognition System Based on Multi-task Convolutional Neural Network and LBP Features. *2019 IEEE International Conference of Intelligent Applied Systems on Engineering* (*ICIASE*), 2019, Pp. 132–135. DOI: 10.1109/ICIASE45644.2019.9074104

3. **Zhao Y., Yang C., Wang Y., Cai J., Xue Y.** Face Recognition for Embedded System Based on Optimized Triplet Loss Neural Network. *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering* (*AEMCSE*), 2020, Pp. 260–263. DOI: 10.1109/AEMCSE50948.2020.00063

4. **Bahnsen F.H., Kaiser J., Fey G.** Designing Recurrent Neural Networks for Monitoring Embedded Devices. *2021 IEEE European Test Symposium* (*ETS*), 2021, Pp. 1–4. DOI: 10.1109/ETS50041.2021.9465460

5. **Chen L., Xiong X., Liu J.** A Survey of Intelligent Chip Design Research Based on Spiking Neural Networks. *IEEE Access*, 2022, Vol. 10, Pp. 89663–89686. DOI: 10.1109/ACCESS.2022.3200454

6. **Zhang M., Gu Z., Pan G.** A Survey of Neuromorphic Computing Based on Spiking Neural Networks. *Chinese Journal of Electronics*, 2018, Vol. 27, No. 4, Pp. 667–674. DOI: 10.1049/cje.2018.05.006

7. **Shinde R.K., Shinde K.D., Mane P.B., Mehta H.** Bio-Inspired Spiking Neural Networks for Real-Time Biomedical Signal Processing. *2025 International Conference on Information, Implementation, and Innovation in Technology* (*I2ITCON*), 2025, Pp. 1–5. DOI: 10.1109/I2ITCON65200.2025.11208866

8. **Chen Y.-N., Chen Y.-H.** Real-Time ECG Recognition Using Convolutional Spiking Neural Networks with Sparse Encoding for Arrhythmia Detection. *2025 IEEE 14th Global Conference on Consumer Electronics* (*GCCE*), 2025, Pp. 380–381. DOI: 10.1109/GCCE65946.2025.11275376

9. **SalimiKia A., Mozaffari S., Ahmadi M., Alirezaee S.** Energy-Efficient Spike Encoding for Implementing Spiking Neural Networks on FPGA. *2025 IEEE Canadian Conference on Electrical and Computer Engineering* (*CCECE*), 2025, Pp. 502–506. DOI: 10.1109/CCECE64018.2025.11364354

10. **Deeksha, Krishnan A., Chandrachoodan N., Nambiar M.** Low-Power FPGA Implementation of Spiking Neural Networks with Optimized Event-Based Encoding. *2025 IEEE 7th International Conference on Emerging Electronics* (*ICEE*), 2025, Pp. 1–4. DOI: 10.1109/ICEE67165.2025.11409777

11. **Chu H., Yan Y., Gan L., Jia H., Qian L., Huan Y.** A Neuromorphic Processing System with Spike-Driven SNN Processor for Wearable ECG Classification. *IEEE Transactions on Biomedical Circuits and Systems*, 2022, Vol. 16, No. 4, Pp. 511–523. DOI: 10.1109/TBCAS.2022.3189364

12. **Davies M., Srinivasa N., Lin T.-H., Chinya G., Cao Y., Choday S.H.** Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*, 2018, Vol. 38, No. 1, Pp. 82–99. DOI: 10.1109/MM.2018.112130359

13. **Ochs M., Dietl M., Brederlow R.** A Circuit Concept for Energy-Efficient Spiking Neural Network Systems with a FOM of 86.9fJ/SOP. *2024 IEEE Biomedical Circuits and Systems Conference* (*BioCAS*), 2024, Pp. 1–5. DOI: 10.1109/BioCAS61083.2024.10798224

14. **Polidori E., Camisa G., Mesri A., Ferrari G., Polidori C., Mastella M.** Experimental validation of an analog spiking neural network with STDP learning rule in CMOS technology. *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering* (*MetroXRAINE*), 2022, Pp. 187–192. DOI: 10.1109/MetroXRAINE54828.2022.9967583

15. **Moriya S., Yamamoto H., Sato S., Yuminaka Y., Horio Y., Madrenas J.** A Fully Analog CMOS Implementation of a Two-variable Spiking Neuron in the Subthreshold Region and its Network Operation. *2022 International Joint Conference on Neural Networks* (*IJCNN*), 2022, Pp. 1–7. DOI: 10.1109/IJCNN55064.2022.9891920

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Danil I. Skrebenkov
Скребенков Данил Игоревич
E-mail: skrebenkov_di@spbstu.ru

Dmitry O. Budanov
Буданов Дмитрий Олегович
E-mail: budanov_do@spbstu.ru
ORCID: https://orcid.org/0000-0002-5925-5628

Submitted: 30.04.2026; Approved: 04.06.2026; Accepted: 13.06.2026.
Поступила: 30.04.2026; Одобрена: 04.06.2026; Принята: 13.06.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 16–30.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 16–30.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19202
UDC 004.8

TFCA-GRU: A TEMPORAL AND FEATURE COOPERATIVE ATTENTION GRU FOR REMAINING USEFUL LIFE PREDICTION OF BEARINGS ACROSS CONDITIONS

C. Guo ✉, *V.V. Potekhin*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ chenxiguo.academic@gmail.com

Abstract. Accurate prediction of the remaining useful life (RUL) of rolling bearings under different operating conditions is essential to enable proactive maintenance of industrial systems. To address the limitations of traditional recursive models in capturing complex time dependencies and distinguishing sensor contributions, we propose a temporal and feature cooperative attention-based gated recursive unit (TFCA-GRU). By integrating temporal and feature channel attention mechanisms, the model adaptively focuses on degradation-related time steps and key sensor features to build more robust RUL models from multidimensional time-frequency fusion data. Therefore, the TFCA-GRU enables sensitive extraction of critical features from long-term dependencies and local feature information. The validity of the proposed TFCA-GRU is verified using the bearing dataset from the XJTU-SY and the IEEE PHM 2012 Prognostic Challenge. This paper also provides the model's intrinsic interpretability through structured non-uniform attentional learning across time and feature dimensions. This reveals TFCA-GRU's selective attention to information time intervals and feature dimensions, which improves the transparency of the model. Experimental results show that the proposed TFCA-GRU performs better than some existing prediction methods in bearing RUL prediction across operating conditions.

Keywords: remaining useful life, deep learning, attention mechanism, gated recurrent unit, time-frequency feature fusion, model interpretability

Acknowledgements: The research was financially supported by the China Scholarship Council (CSC.202309810002).

Citation: Guo C., Potekhin V.V. TFCA-GRU: A temporal and feature cooperative attention GRU for remaining useful life prediction of bearings across conditions. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 16–30. DOI: 10.18721/JCSTCS.19202

© Guo C., Potekhin V.V., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19202
УДК 004.8

TFCA-GRU: АЛГОРИТМ GRU С ВРЕМЕННЫМ И ХАРАКТЕРИСТИЧЕСКИМ КООПЕРАТИВНЫМ ВНИМАНИЕМ ДЛЯ ПРОГНОЗИРОВАНИЯ ОСТАТОЧНОГО СРОКА СЛУЖБЫ ПОДШИПНИКОВ В РАЗЛИЧНЫХ УСЛОВИЯХ

Ч. Го ✉ , *В.В. Потехин*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ chenxiguo.academic@gmail.com

Аннотация. Точное прогнозирование остаточного ресурса (Remaining Useful Life, RUL) подшипников качения в различных условиях эксплуатации является ключевым элементом для организации проактивного обслуживания промышленных систем. Для преодоления ограничений традиционных рекурсивных моделей, которые не всегда способны улавливать сложные временные зависимости и различать вклад отдельных датчиков, мы предлагаем рекуррентный блок с механизмом совместного внимания по временной и признаковой размерностям (TFCA-GRU). Механизм TFCA-GRU интегрирует внимание к временной последовательности и к каналам признаков, что позволяет модели адаптивно фокусироваться на ключевых этапах временной динамики, связанных с деградацией, а также на наиболее информативных сенсорных признаках. Это, в свою очередь, позволяет строить более устойчивые модели RUL на основе многомерных частотно-временных данных. Таким образом, TFCA-GRU обеспечивает более чувствительное извлечение критических признаков из долговременных зависимостей и локальной признаковой информации. Модель была верифицирована на наборах данных XJTU-SY и IEEE PHM 2012 Prognostic Challenge. В работе также подчеркивается внутренняя интерпретируемость модели, которая достигается за счет структурированного неравномерного обучения внимания по временной и признаковой размерностям. Это позволяет раскрыть избирательное внимание TFCA-GRU к информативным временным интервалам и признаковым размерностям, что в свою очередь повышает прозрачность модели. Экспериментальные результаты показывают, что предложенная модель TFCA-GRU превосходит некоторые существующие методы прогнозирования RUL подшипников в различных условиях эксплуатации, что делает ее перспективной для практического применения в промышленных системах.

Ключевые слова: остаточный ресурс, глубокое обучение, механизм внимания, рекуррентный блок, слияние временных и частотных характеристик, интерпретируемость модели

Финансирование: Исследование выполнено при финансовой поддержке Китайского государственного комитета по стипендиям (CSC.202309810002).

Для цитирования: Guo C., Potekhin V.V. TFCA-GRU: A temporal and feature cooperative attention GRU for remaining useful life prediction of bearings across conditions // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 16–30. DOI: 10.18721/JCSTCS.19202

Introduction

Mechanical systems are the foundation of modern industry and accurate prediction of the Remaining Useful Life (RUL) of critical components is essential to ensure reliability and reduce maintenance costs [1, 2]. As a core element of Prognostics and Health Management (PHM), RUL prediction is particularly important for rolling bearings, which are prone to failure and widely used in rotating machinery. Effective RUL estimation enables timely maintenance, minimizes downtime and improves the efficiency and safety of the entire system [3].

© Го Ч., Потехин В.В., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

As the requirements for equipment reliability and operational safety continue to increase, various advanced techniques have been applied to the RUL prediction of rolling bearings. Existing methods are broadly categorized into model-based and data-driven approaches. Model-based approaches rely on physical or mathematical models to simulate degradation behavior, but their applicability in the real world is limited due to the complexity of the failure mechanisms and the need for extensive *a priori* knowledge. In contrast, data-driven approaches utilize sensor data and machine learning algorithms such as support vector regression, random forests and neural networks to learn degradation patterns without the need for explicit physical models. These methods offer greater adaptability and scalability for industrial applications, making them more suitable for complex nonlinear systems [4, 5].

With the rapid development of AI, Deep Learning (DL) techniques hold great promise in the field of PHM, especially in the RUL prediction of rolling bearing [6]. These methods, including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU), can learn degradation patterns directly from raw sensor signals without the need for manual feature extraction. Various strategies have been proposed to improve prediction accuracy and robustness [7]. For example, in [8], a multiscale variational Gaussian self-attention (MVGA) mechanism was developed, that is integrated with a Depth-Wise Sparse Attention Long Short-Term Memory (LSTD) to enhance temporal modelling and mitigate feature confounding through depth propagation. In [9], the focus is on the construction of effective Health Indicators (HIs) via an unsupervised approach by employing a multiscale convolutional LSTM auto-encoder that combines support vector regression and CNN-BiLSTM networks, resulting in improved RUL prediction without the need for manual preconditioning. In [10], the problem of uncertainty modelling and noisy signal interference with the Deep Residual Temporal-Spatial Network (DRTS-Net) was addressed, which integrates attention and soft thresholding to enhance interpretability and probabilistic health estimation.

GRU networks are able to capture long-term temporal dependencies while reducing computational complexity, and thus have become a widely adopted architecture in time series modelling. GRUs are more compact, converge faster, and have a smaller number of parameters than traditional LSTM networks, making them well suited for industrial applications that require real-time and resource-efficient RUL prediction. In recent years, various GRU-based models have been developed to improve the performance of bearing RUL estimation. For example, in [11], a Multiscale Efficient CNN (MSECNN) is proposed, incorporating bidirectional GRUs to exploit spatial and temporal degradation features. They implemented image-based feature extraction using the Gramian Angular Difference Field (GADF) transform to improve the prediction performance under stable conditions. Similarly, in [12], a multi-scale gated CNN-GRU architecture is introduced and used a transformer encoder to improve long-range dependency modelling. In [13], stacked GRUs integrating Dynamic Time Warping (DTW) based attentional mechanisms and Bayesian inference are employed to achieve prediction accuracy and uncertainty estimation. Together, these approaches demonstrate the versatility and potential of GRU-based frameworks for RUL prediction tasks.

Although effective, existing GRU-based models face several significant limitations. Firstly, many models rely on signal transformation or manual feature extraction (e.g., GADF coding or engineering health metrics), which can introduce information loss and increase preprocessing complexity. Secondly, while GRUs can capture temporal dependencies, they lack the intrinsic ability to distinguish which time steps or sensor channels are most informative about degradation progress. While attentional mechanisms have been employed to address this issue, most existing implementations deal with the time and feature dimensions separately or asymmetrically, failing to jointly model their interactions. Moreover, architectures that combine GRUs with complex modules such as transformers or Bayesian estimators typically incur significant computational overheads, which limits their utility in real-world deployments under different operating conditions.

To address these challenges, we propose a novel architecture – Temporal and Feature Cooperative Attention GRU (TFCA-GRU) – which integrates attention mechanisms along the time and feature axes into a unified, lightweight GRU-based framework. This design enables the model to dynamically emphasize critical time steps and degradation-sensitive sensor dimensions, thereby improving prediction accuracy and interpretability under different operating conditions without sacrificing computational efficiency. The main contributions of this work are as follows:

1. A novel time and feature cooperative attention-based GRU framework (TFCA-GRU) is proposed for predicting the RUL of bearings under different operating conditions. The model effectively captures long-range time dependencies and critical sensor interactions through joint attention modelling.

2. A Multi-Head Temporal Attention Mechanism (MTAM) is designed to dynamically highlight the time steps associated with degradation, thereby improving the model's ability to identify critical stages in the degradation process. In addition, a feature-focused module is introduced for adaptively re-weighting input features, thus allowing the model to focus on condition-sensitive signals while maintaining a lightweight structure.

3. Experiments conducted on the XJTU-SY and IEEE PHM 2012 Prognostic Challenge datasets demonstrate the effectiveness and generality of the proposed method. The model shows good prediction accuracy and robustness across operating conditions.

4. The interpretability of the proposed model is verified through attention visualization, which demonstrates its ability to selectively focus on the information time interval and feature dimensions, thus improving the transparency of the model.

Proposed methods

TFCA-GRU framework for bearing RUL

The overall system framework of rolling bearing RUL based on TFCA-GRU is shown in Fig. 1. The system contains four main components, which are: equipment data acquisition, data preprocessing, TFCA-GRU model, and RUL prediction.

GRUs

LSTM networks were introduced to alleviate the gradient vanishing and gradient explosion problems faced by traditional RNNs when processing long sequence-dependent information. LSTMs enhance the capture of long dependencies by designing forgetting gates, input gates, and output gates to selectively store and update time-series information. However, with the increasing demand for network efficiency and model simplicity, GRU has emerged as an improved version of LSTM.

GRU achieves efficient management of information flow by combining the forgetting and input gates of LSTM into a single update gate and introducing a reset gate. This improvement not only reduces the number of network parameters, but also reduces the model complexity and effectively avoids the risk of overfitting. Meanwhile, GRU maintains the sensitivity to information dependence on long series modelling, which further improves the performance of time series prediction.

The core structure of GRU consists of four main states:

1) reset gate r_t,
2) update gate z_t,
3) candidate hidden state $\tilde{h}_t$,
4) output hidden state h_t.

The parameter update equations of GRU model are as follows:

$$r_t = \sigma\left(W_r X_t + U_r h_{t-1} + b_r\right); \tag{1}$$

$$z_t = \sigma\left(W_z X_t + U_z h_{t-1} + b_z\right); \tag{2}$$

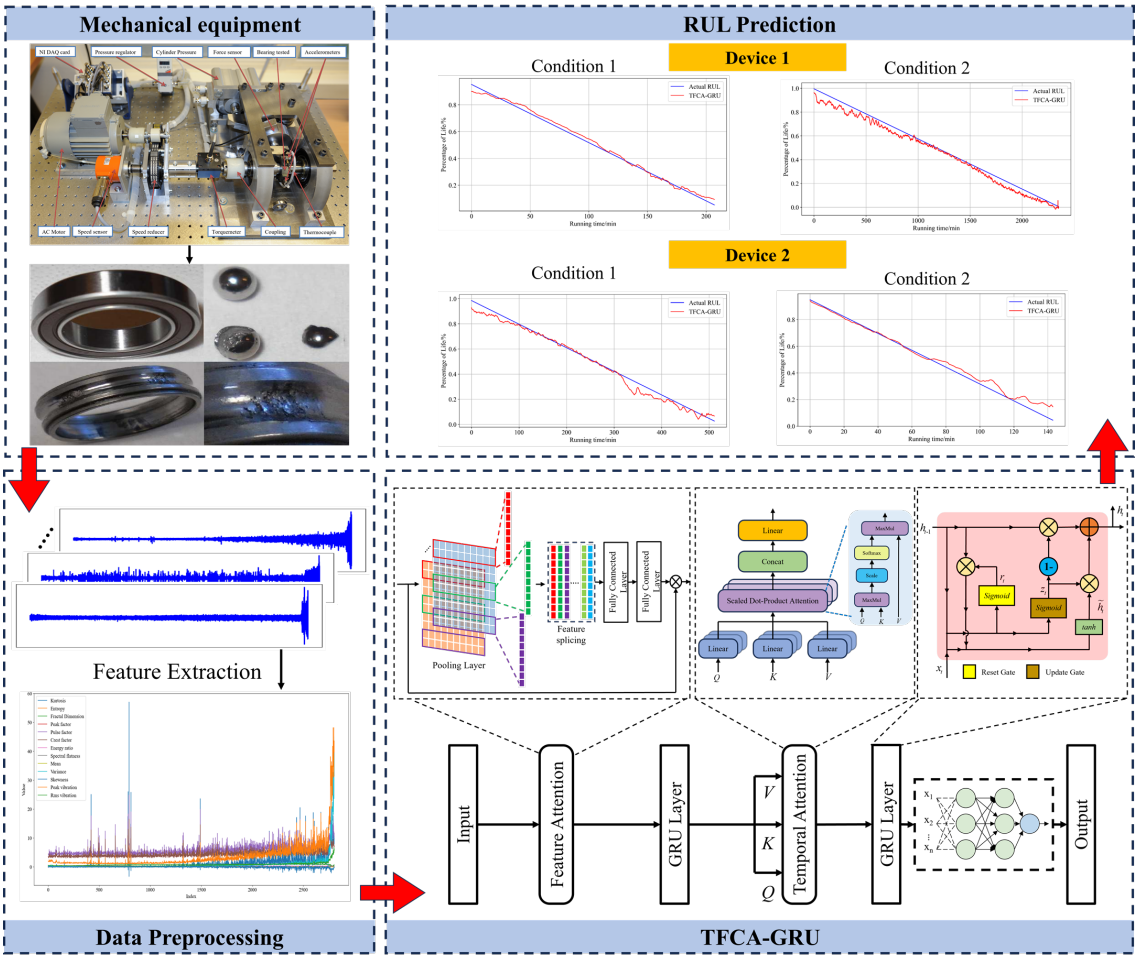


Fig. 1. System framework of the rolling bearing RUL based on TFCA-GRU

$$h_t^* = \tanh\left(WX_t + r_t \odot Uh_{t-1} + b\right); \tag{3}$$

$$h_t = \left(1 - z\right) \odot h_t^* + z_t \odot h_{t-1}, \tag{4}$$

where $\sigma \in (0,1)$ is the activation function, $\tanh \in (-1,1)$ is the hyperbolic tangent function, W and U are the weight matrices, b is the bias term, and $\odot$ is the element-level multiplication.

MTAM

MTAM aims to simulate human's ability to selectively pay attention to time-series information, so that the model can dynamically allocate attention weights, focus on important time steps, and enhance the ability to capture long-time dependencies. Traditional RNNs are limited to a fixed time window, making it difficult to adequately model long-term dependent information. With the introduction of the temporal attention mechanism, however, the model is no longer limited to sequential traversal, but models global associations for all time steps, which substantially improves the flexibility and accuracy of information capture.

The core idea of the temporal attention mechanism designed in this paper is to map the input sequence to three vectors: query, key and value, – and calculate the attention weights through the interaction between these three to determine the degree of attention of the model to different time steps [14].

Suppose the input time sequence is $X = [x_1, x_2, ..., x_T]$, where $x_t \in \mathbb{R}^d$ is the feature vector of the t-th time step, T is the length of the sequence, and d is the feature dimension. In the temporal attention mechanism, the input is first linearly mapped to generate a query vector, a key vector and a value vector. The transformation process of the three vectors is as follows:

$$Q = XW_Q + b_Q; \tag{5}$$

$$K = XW_K + b_K; \tag{6}$$

$$V = XW_V + b_V, \tag{7}$$

where W_Q, W_K and W_V are the weight matrices after linear mapping, b_Q, b_K and b_V are the bias vectors.

The model then calculates the similarity weights of the query and key for each time step using scaled dot product, and the attention weight matrix A is weighted and summed with the value vector V to obtain the final attention representation A:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \tag{8}$$

where A is the attention weight of each time step to the other time steps, and $\sqrt{d}$ is a scaling factor to prevent the gradient from disappearing.

In order to enhance the model's ability to capture multi-level attention subspace information, transformer introduces the multi-head attention mechanism. The advantage of the latter is that it allows model to parallelize the modeling of elements in the sequence from multiple perspectives, and different attention heads can focus on different feature patterns of the input sequence, thus mining the contextual information in a more comprehensive way. The computational process can be represented as follows:

$$\text{head}_j = \text{Attention}\left(QW_j^Q, KW_j^K, VW_j^V\right); \tag{9}$$

$$\text{MultiHead}\left(Q,K,V\right) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)W^H, \tag{10}$$

where $j = 1, 2, ..., h$ and j is the number of heads.

Feature attention mechanism

In multidimensional time series data, there are often important correlations and differences between multiple features. Traditional time series modelling approaches pay fixed attention to individual features, making it difficult to capture the important impact of key features on the forecasting task. In order to better portray these differences, this paper introduces the Feature Attention Mechanism (FAM), which aims to assign higher weights to important features to enhance the model's ability to perceive critical information.

Suppose the input time series is denoted as $X = [x_1, x_2, ..., x_T]$. For any time-step t, the FAM first performs average pooling along the time dimension for all time-steps to capture the global feature performance. The computational expression for this process is:

$$\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t, \tag{11}$$

where $\overline{X} \in \mathbb{R}^{B \times F}$ is the global average representation of each feature over the whole time series. In this paper, we extract a total of 13 dimensional features, so F is 13. Then, the feature dimensions are compressed by nonlinear mapping and the intermediate feature Z is reconstructed using feedforward network to obtain the attention weight A. The computation process is as follows:

$$Z = \mathrm{ReLU}\left(W_1 \overline{X} + b_1\right); \tag{12}$$

$$A = \sigma\left(W_2 Z + b_2\right), \tag{13}$$

where W_1 and W_2 are the weight matrices, b_1 and b_2 are the bias terms, and σ is the Sigmoid activation function.

Finally, in order to be consistent with the dimension of the input vectors, the attention weights A are dimensionally extended and the attention weighting operation is performed with the original features, and the output of the FAM is computed as follows:

$$X' = X \odot A', \tag{14}$$

where A' is the expanded attention weight, $\odot$ is the elemental multiplication, and X' is the feature-weighted output.

Prediction process of rolling bearing RUL based on TFCA-GRU

Step 1: The vibration signal is collected by industrial equipment to obtain the original time series data.

Step 2: Based on the feature extraction methods in time and frequency domains, the vibration signals are analyzed and converted into multi-dimensional features to generate the time-frequency domain fusion feature matrix.

Step 3: The processed data set is divided into training set and test set, and the training set is input into the constructed TFCA-GRU network model for multiple iterations optimization, to gradually learn the deterioration law of the equipment and generate the final RUL prediction model.

Step 4: Input the test set into the trained model to complete the accurate prediction of the RUL of rolling bearings.

Experiments and results analysis

The model modelling and experimental tests in this paper are developed based on the PyTorch DL framework using Python version 3.11. The development environment is PyCharm 2023. The experimental hardware configuration is Intel Xeon(R) Platinum 8352 CPU and NVIDIA GeForce RTX 4090 GPU.

Dataset and data preprocessing

In this article, the IEEE PHM 2012 Prognostic Challenge dataset [15] and the XJTU-SY dataset [16] will be used for experimental testing. The working conditions of the dataset and the division of the dataset are shown in Table 1.

The IEEE PHM 2012 Prognostic Challenge dataset was collected from the PRONOSTIA platform, which has been widely used in many equipment residual life prediction studies. The platform consists of a rotary drive system, radial loading device and measurement system, and is capable of simulating the degradation process of rolling bearings under real operating conditions. The vibration signals are collected at a sampling frequency of 25.6 kHz, and data segments of 0.1 second duration are recorded every 10 seconds, which is equivalent to 2560 sampling points per segment. When the amplitude of the monitored vibration exceeds 20 G, the experiment is judged as bearing failure and terminated. In this paper, the training set is used in strict accordance with the official classification scheme, and the rest of the bearing data is used as the test set.

Table 1

Running conditions and experimental setup for the dataset

Datasets	Operating conditions	Training dataset	Test dataset
IEEE PHM 2012 Prognostic Challenge	1800 rpm, 4 kN	Bearing1_1 Bearing1_2	Bearing1_3, Bearing1_4, Bearing1_5, Bearing1_6, Bearing1_7
	1650 rpm, 4.2 kN	Bearing2_1 Bearing2_2	Bearing2_3, Bearing2_4, Bearing2_5, Bearing2_6, Bearing2_7
XJTU-SY	2100 rpm, 12 kN	Bearing1_1 Bearing1_2	Bearing1_3, Bearing1_4, Bearing1_5
	2250rpm, 11 kN	Bearing2_1 Bearing2_2	Bearing2_3, Bearing2_4, Bearing2_5

The XJTU-SY dataset is based on a highly integrated bearing life test platform, which includes a motor system for controlling the rotational speed, a hydraulic control device for loading, and an array of multi-channel vibration sensors for real-time acquisition of vibration signals during operation. The platform supports flexible adjustment of radial load and rotational speed to simulate bearing degradation under different operating conditions. The data is collected at a sampling frequency of 25.6 kHz, with complete samples taken every 1.28 seconds, each containing 32768 consecutive data points, providing high-resolution, long-period vibration sequences for subsequent failure modelling and remaining life prediction.

Furthermore, the 13 time-domain and frequency-domain features were extracted from the vibration signals of the dataset species. The time-domain features included kurtosis, mean, variance, skewness, peak vibration, root-mean-square vibration, pulse factor, crest factor, and peak factor, while the frequency-domain features covered entropy, fractal dimension, energy ratio, and spectral flatness.

Assessment metrics and comparison algorithms

In the field of RUL prediction, the following three methods are commonly used to assess the prediction performance of the model [17, 18]:

1) Root Mean Square Error (RMSE),
2) Mean Absolute Error (MAE),
3) scoring function (Score) and R-Square (R^2) [19].

The three assessments are calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2}; \tag{15}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y_i} - y_i\right|; \tag{16}$$

$$\mathrm{Score} = \begin{cases} \sum_{i=1}^{N_2}\left(e^{-\frac{R^i_{\mathrm{pre}} - R^i_{\mathrm{true}}}{13}} - 1\right), & R^i_{\mathrm{pre}} < R^i_{\mathrm{true}} \\ \sum_{i=1}^{N_2}\left(e^{-\frac{R^i_{\mathrm{pre}} - R^i_{\mathrm{true}}}{10}} - 1\right), & R^i_{\mathrm{pre}} \geq R^i_{\mathrm{true}} \end{cases}; \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2}{\sum_{i=1}^{n}\left(\overline{y_i} - y_i\right)^2}, \tag{18}$$

where n is the number of samples, y_i and $\widehat{y_i}$ are the actual and predicted RUL values at the i-th moment, $\overline{y}$ is the actual RUL mean, and N_2 is the total number of test samples. Generally, lower values of RMSE and MAE indicate better performance of the model. The closer the value of R^2 is to 1 indicates higher prediction accuracy of the model.

In order to demonstrate the superiority of the proposed TFCA-GRU, five representative baseline RUL prediction methods are used in this paper to compare the RUL prediction effectiveness, as well as to compare the evaluation data with the current novel RUL prediction methods. The baseline methods include CNN, LSTM, CNN-LSTM, CNN-Transformer, and the novel methods include MSCNN [20], TCN-SA [21], DCTCN [22], DSCN [23].

Experiment 1: PHM 2012 dataset

The RUL prediction results of the proposed TFCA-GRU with the baseline model are illustrated in Fig. 2. Firstly, from the prediction curves, the proposed model can track the actual RUL degradation trend stably under all four operating conditions, and shows good dynamic response capability in both the initial and termination stages of degradation. In contrast, the baseline approach exhibits different degrees of drift and prediction lag under several scenarios, which is difficult to accurately reflect the system degradation evolution. In the box-and-line plot, TPCA-GRU shows the smallest median prediction error and small distribution variance in all scenarios, and the fewest outliers, which reflects its advantages of prediction stability and robustness.

The evaluation data of the proposed TFCA-GRU with the novel models on the PHM 2012 dataset are shown in Table 2. From the overall trend, TFCA-GRU significantly outperforms the other four models in terms of MAE and RMSE metrics under all data scenarios, showing strong consistency and robustness. In addition, a side-by-side comparison between the models shows that although the SCTCN, DSCN and TCN-SA models perform the best MAE and RMSE metrics in multiple scenarios, they exhibit unstable prediction and large error fluctuations in cross-case conditions.

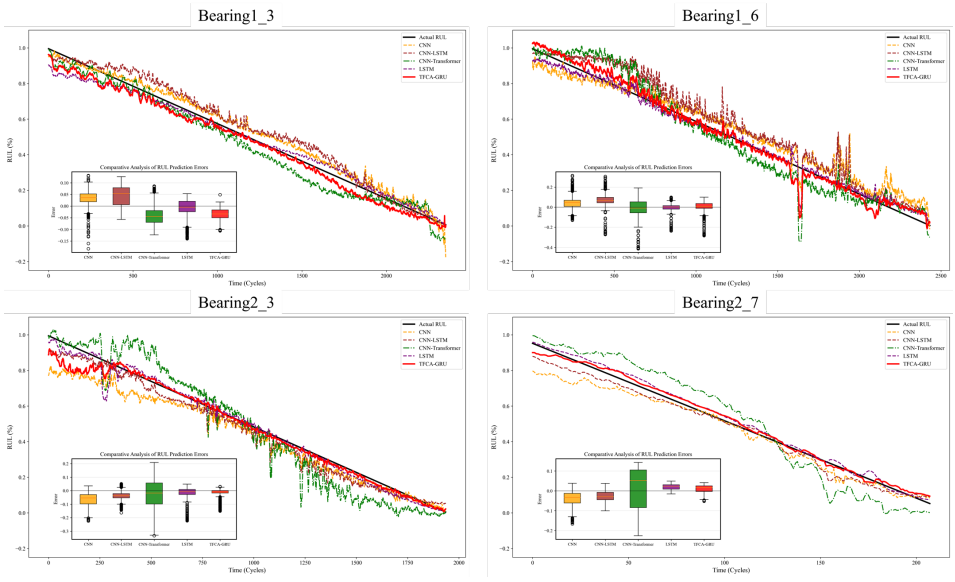


Fig. 2. Comparison of the prediction results for the four working conditions in the PHM 2012 dataset

Table 2

Performance comparison on PHM 2012 dataset

	MSCNN		TCN-SA		SCTCN		DSCN		TFCA-GRU	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
1–3	0.078	0.096	0.066	0.085	0.098	0.121	0.071	0.085	**0.035**	**0.041**
1–4	**0.015**	0.180	0.103	0.139	0.130	0.149	0.084	0.108	0.037	**0.045**
1–5	0.089	0.151	0.086	0.126	0.163	0.222	0.087	0.146	**0.051**	**0.040**
1–6	0.105	0.149	0.084	0.116	0.178	0.221	0.084	0.112	**0.030**	**0.042**
1–7	0.106	0.147	0.089	0.125	0.084	0.109	0.089	0.121	**0.044**	**0.057**
2–3	0.269	0.325	0.191	0.232	0.110	0.216	0.191	0.238	**0.024**	**0.042**
2–4	0.185	0.226	0.091	0.129	0.117	0.145	0.257	0.298	**0.021**	**0.030**
2–5	0.137	0.176	0.192	0.222	0.189	0.221	0.131	0.171	**0.034**	**0.074**
2–6	0.103	0.126	0.08	0.098	0.092	0.115	0.101	0.130	**0.031**	**0.047**
2–7	0.241	0.351	0.196	0.230	0.167	0.195	0.216	0.255	**0.019**	**0.022**

Experiment 2: XJTU-SY dataset

The RUL prediction result plots of the proposed TFCA-GRU with the baseline model are illustrated in Fig. 3. From the time-series prediction curves, traditional deep models such as CNN and LSTM have obvious lagging or advancing phenomena, especially in the early and late stages of degradation, with the largest prediction deviation. The proposed model is able to closely track the degradation trend of the real RUL decreasing over time, especially in Bearing1_3 and Bearing2_5, where its prediction curves almost overlap with the real curves, indicating its strong ability to capture the degradation features. As seen from the boxplots, the proposed model shows the smallest median error, narrower error intervals, and very few outliers in all four subplots, which indicates that its predictions are not only accurate but also stable.

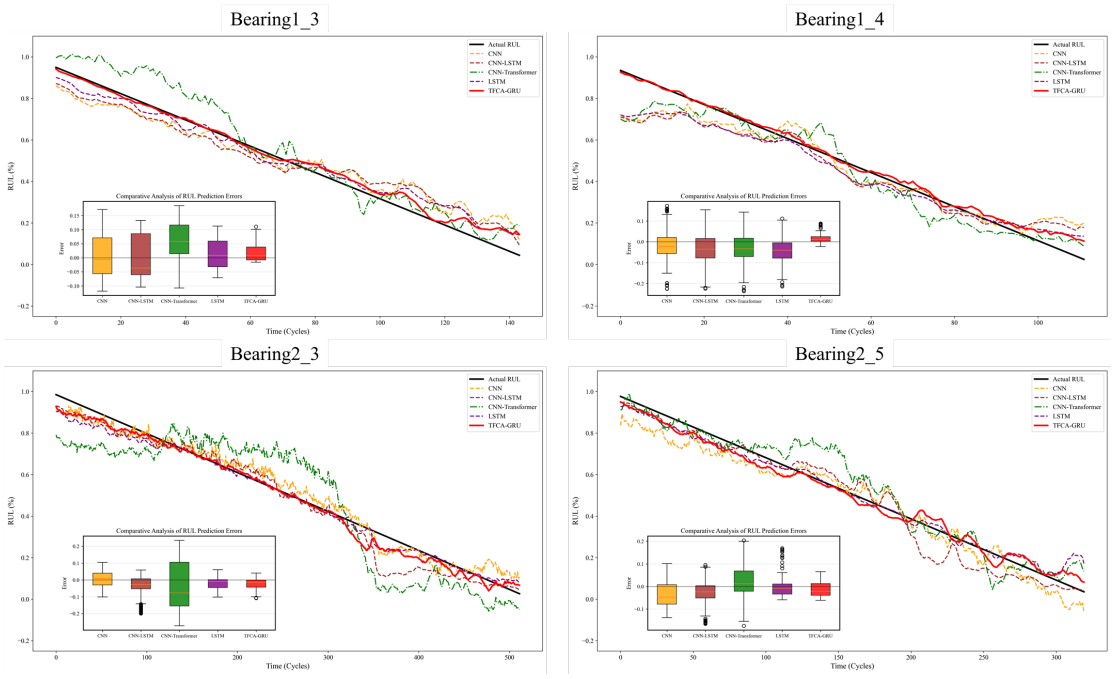


Fig. 3. Comparison of the prediction results of the four operating conditions on XJTU-SY dataset

The evaluation data of the proposed TFCA-GRU and the new model on the XJTU-SY dataset are presented in Table 3. It can be seen that the MAE and RMSE values of TFCA-GRU are significantly better than the other models in both working conditions. Compared with the other models, TCN-SA and DSCN, although better in some indicators, are difficult to maintain a stable advantage under multiple operating conditions.

Table 3

Performance comparison on XJTU-SY dataset

	MSCNN		TCN-SA		SCTCN		DSCN		TFCA-GRU	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
1–3	0.227	0.262	0.120	0.139	0.114	0.133	0.092	**0.104**	**0.083**	0.110
2–3	0.086	0.104	0.092	0.118	0.160	0.191	0.094	0.118	**0.063**	**0.086**

Ablation experiment

The ablation models are set up as GRU, TA-GRU with the introduction of temporal attention, FA-GRU with the introduction of feature attention, and TFCA-GRU.

The prediction curves of the ablation model in XJTU-SY Bearing2_5 are illustrated in Fig. 4. It can be seen that the prediction curve of TFCA-GRU basically coincides with the real RUL, which fully demonstrates the effectiveness of the joint time-feature modelling. Compared with other models, TFCA-GRU is able to extract the key dependencies in the degraded signals more comprehensively, thus achieving higher accuracy and more robust life prediction.

The comparison of the comprehensive evaluation of the models is demonstrated in Table 4. GRU performs relatively poorly on all metrics, showing its limited predictive performance when no attention mechanism is introduced. In contrast, TFCA-GRU has the best overall performance on all metrics. Therefore, this justifies TFCA-GRU.

Table 4

Performance comparison of ablation models

Ablation models	R2	RMSE	MAE	Score
GRU	0.954	0.058	0.148	1.404
TA-GRU	0.987	0.037	0.033	0.885
FA-GRU	0.971	0.046	0.037	1.109
TFCA-GRU	**0.991**	**0.031**	**0.026**	**0.779**

Model interpretability analysis

First, we compare and analyze the effect of its temporal attention on the model's time-awareness ability, and Fig. 5 demonstrates the heat map of attention in the temporal dimension. TFCA-GRU presents obvious non-uniformity, and the model is able to dynamically distribute the degree of attention at different time steps, which significantly enhances the ability of identifying the key degradation phases; whereas the GRU model has a more uniform distribution of attention, and lacks the ability of efficiently capturing the important timing information is effectively captured.

Meanwhile, Fig. 6 demonstrates the distribution of the model's attentional weights on feature dimensions. GRU has a relatively even distribution of weights across feature dimensions, and the model lacks differentiation of the importance of input features. TFCA-GRU attentional distributions show significant differences, with some feature dimensions (e.g., indexes 0, 10, etc.) obtaining higher

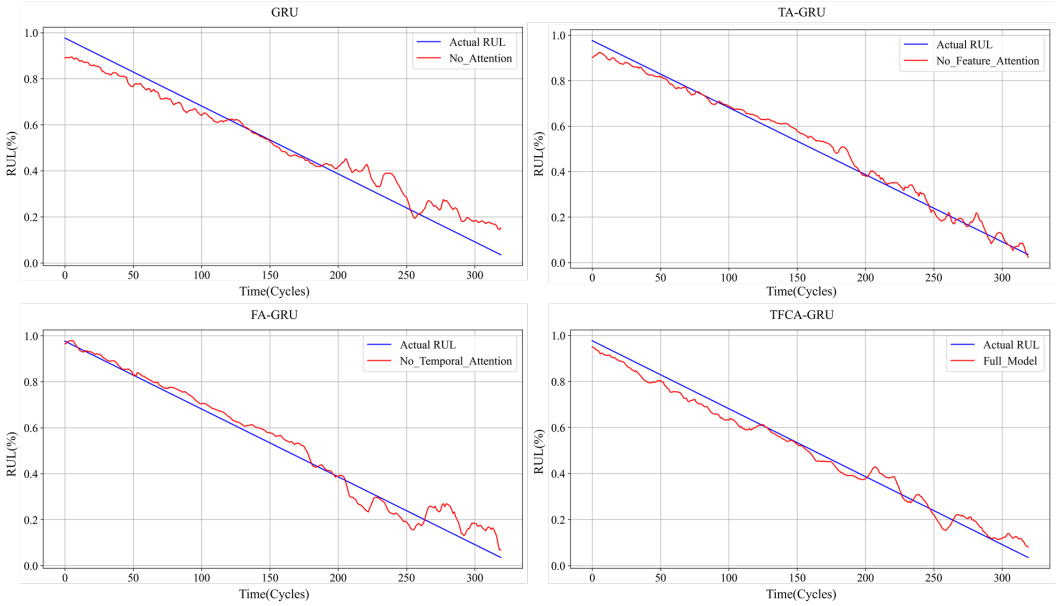


Fig. 4. Comparison of predictions from ablation models

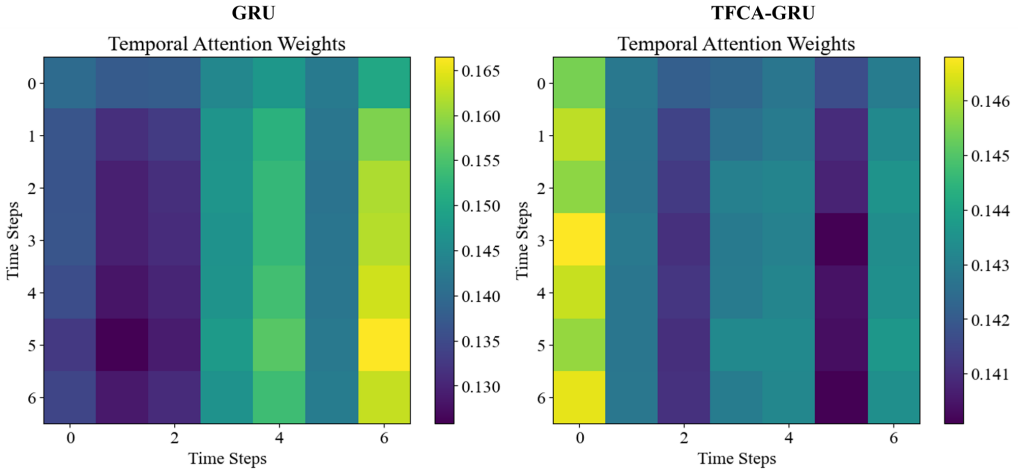


Fig. 5. Heat map of temporal attention

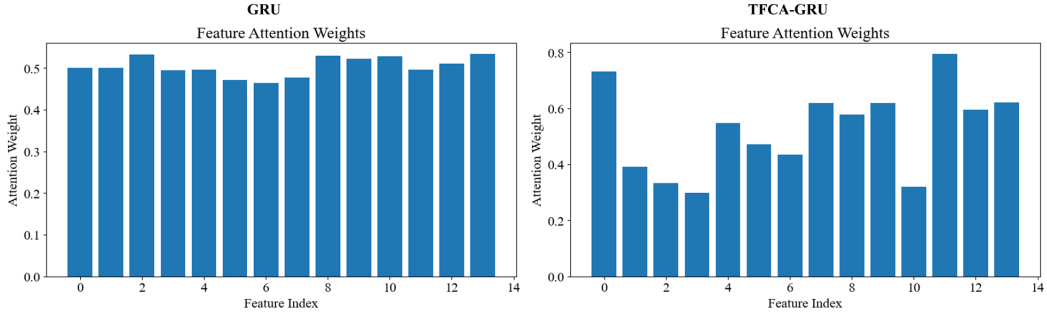


Fig. 6. Distribution of feature attention weights

weights, suggesting that the model is capable of adaptively identifying and reinforcing the features that are most relevant to the trend of degradation.

In this study, we proposed TFCA-GRU, a temporal and feature cooperative attention-based gated recurrent unit framework for the prediction of the rolling bearings RUL under variable operating conditions. The model addresses two core challenges in industrial prognostics: the difficulty of capturing long-term temporal dependencies and the inability of conventional models to distinguish between informative and irrelevant sensor features.

Conclusions and further work

By incorporating MTAM and FAM into the GRU architecture, TFCA-GRU is capable of dynamically focusing on degradation-critical time steps and sensor dimensions. This joint attention design enhances the model's ability to learn representative patterns from multidimensional time-frequency fusion data while preserving computational efficiency. Unlike traditional GRU-based or hybrid deep models that rely heavily on handcrafted preprocessing or suffer from scalability issues, TFCA-GRU offers a unified and interpretable solution.

Comprehensive experiments on two public datasets – IEEE PHM 2012 Prognostic Challenge and XJTU-SY – demonstrated that TFCA-GRU consistently outperforms baseline and state-of-the-art models in terms of prediction accuracy, robustness, and generalization across working conditions. Ablation studies confirmed the effectiveness of each attention component, and visualization of attention distributions further illustrated the model's capacity to highlight key temporal intervals and dominant features, thereby improving transparency and trustworthiness in critical applications.

Overall, TFCA-GRU provides a practical and extensible framework for data-driven RUL prediction under complex and non-stationary degradation scenarios. Future research will focus on further reducing the model's computational overhead, exploring domain adaptation techniques to improve cross-machine transferability, and extending the framework to other types of mechanical components and multi-sensor fusion settings.

REFERENCES

1. **Sai Van Cuong, Shcherbakov M.V.** Architecture of predictive maintenance system of complex multiobject systems in Industry 4.0 concept. *Software & Systems*, 2020, Vol. 33, No. 2, Pp. 186–194. DOI: 10.15827/0236-235X.130.186-194

2. **Cui L., Shen Q., Xiao Y., Liu D., Wang H.** Sparse graph structure fusion convolutional network for machinery remaining useful life prediction. *Reliability Engineering & System Safety*, 2025, Vol. 254A, Art. no. 110592. DOI: 10.1016/j.ress.2024.110592

3. **Dli M.I., Puchkov A.Yu., Lobaneva E.I.** Method for estimating the time of useful use of equipment based on neural networks. *Bulletin of the Saint Petersburg State Institute of Technology* (*Technical University*), 2021, Vol. 85, No. 59, Pp. 107–112. DOI: 10.36807/1998-9849-2021-59-85-107-112

4. **Zhang J., Tian J., Yan P., Wu S., Luo H., Yin S.** Multi-hop graph pooling adversarial network for cross-domain remaining useful life prediction: A distributed federated learning perspective. *Reliability Engineering & System Safety*, 2024, Vol. 244, Art. no. 109950. DOI: 10.1016/j.ress.2024.109950

5. **Xu Z., Li C., Yang Y.** Fault diagnosis of rolling bearings using an Improved Multi-Scale Convolutional Neural Network with Feature Attention mechanism. *ISA Transactions*, 2021, Vol. 110, Pp. 379–393. DOI: 10.1016/j.isatra.2020.10.054

6. **Liu H., Sun Y., Ding W., Wu H., Zhang H.** Enhancing non-stationary feature learning for remaining useful life prediction of aero-engine under multiple operating conditions. *Measurement*, 2024, Vol. 227, Art. no. 114242. DOI: 10.1016/j.measurement.2024.114242

7. **Yin Y., Tian J., Liu X.** Remaining useful life prediction based on parallel multi-scale feature fusion network. *Journal of Intelligent Manufacturing*, 2024, Vol. 36, Pp. 3111–3127. DOI: 10.1007/s10845-024-02399-y

8. **Fan W., Chang Z., Xu S., Fan Z., Yuan Y.** RUL Prediction of Rolling Bearings with MVGA-DSAL: A Multiscale Variant Gaussian Attention Model with Depth-Wise Sparse Attention LSTM. *IEEE Transactions on Instrumentation and Measurement*, 2025, Vol. 74, Art no. 3523111. DOI: 10.1109/TIM.2025.3551998

9. **Qu Y., Fu S., Yong M., Tian J., Lv Z., Li R.** Health Indicator Construction and Remaining Useful Life Prediction Based on MSC-LSTM-AE Model for Working Bearings. *IEEE Sensors Journal*, 2025, Vol. 25, No. 9, Pp. 15525–15535. DOI: 10.1109/JSEN.2025.3548675

10. **Yang T., Li G., Huo J., Li X., Zhou X., Han Q.** The DRTS-Net Framework: Enhanced Probabilistic Health Perception for Wind Turbine Bearings Under Complex Operating Conditions with Improved Interpretability. *IEEE Transactions on Instrumentation and Measurement*, 2025, Vol. 74, Art no. 3517412. DOI: 10.1109/TIM.2025.3545504

11. **Ma P., Li G., Zhang H., Wang C., Li X.** Prediction of Remaining Useful Life of Rolling Bearings Based on Multiscale Efficient Channel Attention CNN and Bidirectional GRU. *IEEE Transactions on Instrumentation and Measurement*, 2024, Vol. 73, Art no. 2508413. DOI: 10.1109/TIM.2023.3347787

12. **Guo D., Cao Z., Fu H., Li Z.** Remaining Useful Life Estimation for Rolling Bearings Using MSGCNN-TR. *IEEE Sensors Journal*, 2022, Vol. 22, No. 24, Pp. 24333–24343. DOI: 10.1109/JSEN.2022.3221753

13. **Que Z., Jin X., Xu Z.** Remaining Useful Life Prediction for Bearings Based on a Gated Recurrent Unit. *IEEE Transactions on Instrumentation and Measurement*, 2021, Vol. 70, Art no. 3511411. DOI: 10.1109/TIM.2021.3054025

14. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.** Attention is all you need. *arXiv:1706.03762*, 2017. DOI: 10.48550/arXiv.1706.03762

15. **Nectoux P., Gouriveau R., Medjaher K., Ramasso E., Chebel-Morello B. et al.** PRONOSTIA: An experimental platform for bearings accelerated degradation tests. *IEEE International Conference on Prognostics and Health Management* (*PHM'12*), 2012, pp. 1–8.

16. **Wang B., Lei Y., Li N., Li N.** A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, 2020, Vol. 69, No. 1, Pp. 401–412. DOI: 10.1109/TR.2018.2882682

17. **Wang T., Liu H., Guo D., Sun X.-M.** Continual residual reservoir computing for remaining useful life prediction. *IEEE Transactions on Industrial Informatics*, 2024, Vol. 20, No. 1, Pp. 931–940. DOI: 10.1109/TII.2023.3271661

18. **Qin Y., Chen D., Xiang S., Zhu C.** Gated Dual Attention Unit Neural Networks for Remaining Useful Life Prediction of Rolling Bearings. *IEEE Transactions on Industrial Informatics*, 2021, Vol. 17, No. 9, Pp. 6438–6447. DOI: 10.1109/TII.2020.2999442

19. **Babu G.S., Zhao P., Li X.-L.** Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life. *Database Systems for Advanced Applications* (*DASFAA 2016*), 2016, Vol. 9642, Pp. 214–228. DOI: 10.1007/978-3-319-32025-0_14

20. **Zhu J., Chen N., Peng W.** Estimation of Bearing Remaining Useful Life Based on Multiscale Convolutional Neural Network. *IEEE Transactions on Industrial Electronics*, 2019, Vol. 66, No. 4, Pp. 3208–3216. DOI: 10.1109/TIE.2018.2844856

21. **Wang Y., Deng L., Zheng L., Gao R.X.** Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics. *Journal of Manufacturing Systems*, 2021, Vol. 60, Pp. 512–526. DOI: 10.1016/j.jmsy.2021.07.008

22. **Zou H., Zhu J.** DCTCN: Deep Complex Temporal Convolutional Network for Real Time Speech Enhancement. *2021 11th International Conference on Intelligent Control and Information Processing* (*ICICIP*), 2021, Pp. 112–118. DOI: 10.1109/ICICIP53388.2021.9642159

23. **Hu K., Yan A.** Multitarget Robust Deep Stochastic Configuration Network Parameter Modeling Method. *2024 6th International Conference on Industrial Artificial Intelligence* (*IAI*), 2024, Pp. 1–5. DOI: 10.1109/IAI63275.2024.10729968

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Guo Chenxi
Го Чэньси
E-mail: chenxiguo.academic@gmail.com
ORCID: https://orcid.org/0000-0002-1497-2893

Vyacheslav V. Potekhin
Потехин Вячеслав Витальевич
E-mail: Slava.Potekhin@spbstu.ru
ORCID: https://orcid.org/0000-0001-9850-9558

Submitted: 24.05.2025; Approved: 25.05.2026; Accepted: 03.06.2026.
Поступила: 24.05.2025; Одобрена: 25.05.2026; Принята: 03.06.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 31–43.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 31–43.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19203
UDC 004.85

SURVIVAL ANALYSIS USING MAHALANOBIS DISTANCE IN KERNELS OF THE BERAN ESTIMATOR

L.V. Utkin ✉, *A.V. Konstantinov*, *S.R. Kirpichenko*, *N.M. Verbova*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ lev.utkin@gmail.com

Abstract. Survival analysis is critical for modeling time-to-event data across medicine, engineering, and economics. The Beran estimator serves as an efficient nonparametric tool for estimating conditional survival functions under censoring. However, the conventional Beran estimator relies on the Euclidean distance for kernel weighting, a metric that fails to account for heterogeneous scales and correlations among covariates. This limitation can result in suboptimal weighting when features exhibit multicollinearity or varying measurement units. To address this, we propose the M-Beran estimator, a generalized formulation that incorporates the Mahalanobis distance into the kernel smoothing scheme. This modification ensures scale invariance and leverages the underlying correlation structure of the feature space, transforming the kernel into an elliptically symmetric form. We evaluate the proposed method through numerical experiments on real-world survival datasets. Results demonstrate that the M-Beran estimator outperforms the conventional Euclidean-distance-based Beran estimator, offering improved predictive accuracy.

Keywords: machine learning, survival analysis, Beran estimator, Mahalanobis distance, Harrell's C-index, survival function

Acknowledgements: The research was financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment "Development and research of machine learning models for solving fundamental problems of artificial intelligence for the fuel and energy complex" (FSEG-2024-0027).

Citation: Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Verbova N.M., Survival analysis using Mahalanobis distance in Kernels of the Beran estimator. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 31–43. DOI: 10.18721/JCSTCS.19203

© Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Verbova N.M., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19203
УДК 004.85

АНАЛИЗ ВЫЖИВАЕМОСТИ С ИСПОЛЬЗОВАНИЕМ РАССТОЯНИЯ МАХАЛАНОБИСА В ЯДРАХ МОДЕЛИ БЕРАНА

Л.В. Уткин ✉, *А.В. Константинов*, *С.Р. Кирпиченко*, *Н.М. Вербова*

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

✉ lev.utkin@gmail.com

Аннотация. Анализ выживаемости имеет решающее значение для моделирования данных о времени до наступления события в медицине, инженерии и экономике. Оценка Берана служит эффективным непараметрическим инструментом для оценки условных функций выживаемости при цензурированных данных. Однако стандартная оценка Берана основана на евклидовом расстоянии для взвешивания ядер – метрике, которая не учитывает различие в масштабах признаков и корреляции между ними. Эти ограничения могут привести к неоптимальному взвешиванию, когда признаки демонстрируют мультиколлинеарность или различные единицы измерения. Для решения этой проблемы мы предлагаем оценку М-Берана – обобщенную формулировку, которая включает расстояние Махаланобиса в схему сглаживания ядер. Эта модификация обеспечивает инвариантность к масштабу данных и использует лежащую в основе корреляционную структуру пространства признаков, преобразуя ядро в эллиптически симметричную форму. Мы оцениваем предложенный метод с помощью численных экспериментов на реальных наборах данных времен до событий. Результаты показывают, что оценка М-Берана превосходит традиционную оценку Берана на основе евклидова расстояния, обеспечивая более высокую точность предсказаний.

Ключевые слова: машинное обучение, анализ выживаемости, оценка Берана, расстояние Махаланобиса, С-индекс, функция выживаемости

Финансирование: Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка и исследование моделей машинного обучения для решения фундаментальных задач искусственного интеллекта в топливно-энергетическом комплексе» (FSEG-2024-0027).

Для цитирования: Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Survival analysis using Mahalanobis distance in Kernels of the Beran estimator // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 31–43. DOI: 10.18721/JCSTCS.19203

Introduction

Survival analysis, the statistical study of time-to-event data, is essential across diverse fields [1–4]. It guides clinical trials and disease modeling in medicine, predicts system failure in engineering, and assesses churn or turnover in economics, social sciences, and marketing. Characterized by censoring, where events may be unobserved, this data complexity motivates the continued development of robust estimation methods. Among the most fundamental tools for flexible survival analysis is the Beran estimator [5]. Given a set of censored survival times and a vector of covariates, the Beran estimator provides a nonparametric estimate of the conditional survival function (SF) by applying kernel smoothing directly to the observed events and censoring processes. Its appeal lies in its ability to model complex, nonlinear relationships without imposing restrictive parametric assumptions such as proportional hazards [6], making it particularly valuable in applications where the underlying mechanisms are poorly

© Уткин Л.В., Константинов А.В., Кирпиченко С.Р., Вербова Н.М., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

understood or where standard models fail. The estimator's formulation relies on a kernel function that assigns weights to neighboring observations based on their proximity in the feature space, typically employing the Euclidean distance.

Despite its flexibility, the conventional Beran estimator inherits a fundamental limitation from its reliance on the Euclidean metric. In practical applications, covariate vectors often exhibit heterogeneous scales, complex correlations, and multicollinearity – features that the Euclidean distance is ill-equipped to handle. Consider, for instance, a clinical study of cancer survival where covariates include tumor size (measured in millimeters), patient age (in years), and multiple correlated biomarkers; in engineering reliability, correlated sensor measurements from a monitoring system. In each case, the Euclidean metric treats each covariate dimension independently and equally. Moreover, it ignores the underlying correlation structure of the covariate distribution, which can lead to suboptimal weighting: two patients who appear similar in Euclidean terms may, in fact, be statistically dissimilar if they deviate from the dominant correlation patterns of the biomarker panel, potentially compromising the accuracy of survival predictions that inform critical clinical decisions.

A solution to these challenges lies in replacing the Euclidean distance with the Mahalanobis distance [7].

The Mahalanobis distance offers several advantages that are directly relevant to real-world survival analysis applications. First, it is scale-invariant, automatically accounting for differences in measurement units, whether tumor size in millimeters, gene expression on a log scale, or sensor voltages in engineering systems, and eliminating the need for separate bandwidths per covariate. Second, and more importantly, it incorporates the correlation structure of the data, so that distances are measured in terms of statistical fluctuations rather than raw coordinate differences. Observations that lie along directions of high variance are considered closer, while those that deviate against the correlation structure are appropriately penalized. This property is valuable in applications with redundant or highly correlated predictors, such as the analysis of multiple clinical biomarkers, correlated survey responses in social science event history studies, or multi-sensor reliability data.

In the context of kernel-based nonparametric estimation, substituting the Euclidean distance with the Mahalanobis distance transforms the kernel from a product of independent univariate kernels into a single, elliptically symmetric kernel defined on the standardized and decorrelated feature space. This simple modification carries profound implications for applied survival analysis. It can help to maximize the predictive performance of the Beran estimator, such as by maximizing the conditional likelihood of the observed survival times. While the original Beran estimator and its theoretical properties have been extensively studied [5, 8, 9], the systematic incorporation of the Mahalanobis distance into its kernel structure remains relatively underexplored.

The Mahalanobis distance attracts attention in tasks of survival analysis. For example, a nonparametric weighting approach was proposed in [10], which balances the covariates between treatment groups by Kernel Mean Matching (KMM) with the Mahalanobis distance kernels. A linear-mixed effects model and the Cox model for survival time are considered in [11, 12], where the Mahalanobis distance is used to detect outliers in survival data. In [13], a Moving Multivariate Distance (MMD), based on the Mahalanobis distance, was proposed to quantify the variability of the multivariate biomarker profile as a whole from one visit to the next. In [14], flexible machine learning models were evaluated, capable of incorporating interactions and nonlinear effects from high-dimensional data using the Mahalanobis distance to estimate individualized treatment recommendations in trials with time-to-event outcomes. In [15], a method was proposed to match each untreated patient with the nearest treated patient based on the Mahalanobis distance between patient characteristics.

Techniques based on the Mahalanobis distance and applied in different fields of chemometrics such as in multivariate calibration, pattern recognition and process control are explained and discussed in [16]. The most important application of the Mahalanobis distance is related to out-of-distribution

detection and outlier detection [17–20]. Application of the Mahalanobis distance to a k-means clustering algorithm was studied in [21]. In [22], it is proposed using a Mahalanobis distance metric for computing the attention weights to stretch the underlying feature space in directions of high contextual relevance.

However, to the best of our knowledge, at the moment, no approach is known that incorporates the Mahalanobis distance to the Beran estimator. Therefore, following [22], we propose a generalized Beran estimator that incorporates the Mahalanobis distance into its kernel weighting scheme. We called the proposed model as the *M-Beran estimator*. Through real data, it is shown that the M-Beran estimator outperforms the conventional Beran estimator, using the Euclidean distance.

Background

The Mahalanobis Distance

The Mahalanobis distance is a measure of distance that accounts for the scale and correlation structure of multivariate data [7]. Unlike the Euclidean distance, which treats all dimensions independently, the Mahalanobis distance provides a metric that is invariant to affine transformations of the covariate space. Let $\boldsymbol{X} = \left(x_1, \ldots, x_p\right)^T$ be a p-dimensional random vector with mean vector $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{X}\right]$ and positive-definite covariance matrix

$$\boldsymbol{M} = \mathrm{Cov}\left(\boldsymbol{X}\right) = \mathbb{E}\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T\right].$$

For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, the squared Mahalanobis distance is defined as

$$d_M^2\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left(\boldsymbol{x} - \boldsymbol{y}\right)^T \boldsymbol{M}^{-1}\left(\boldsymbol{x} - \boldsymbol{y}\right).$$

When comparing an observation $\boldsymbol{x}$ to the population mean $\boldsymbol{\mu}$, the squared Mahalanobis distance simplifies to

$$d_M^2\left(\boldsymbol{x}, \boldsymbol{\mu}\right) = \left(\boldsymbol{x} - \boldsymbol{\mu}\right)^T \boldsymbol{M}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}\right).$$

The Mahalanobis distance possesses several properties that make it particularly attractive for statistical applications. In practice, the population parameters $\boldsymbol{\mu}$ and $\boldsymbol{M}$ are unknown and can be estimated from data.

Concepts of Survival Analysis

Unlike standard regression settings, survival data are characterized by the presence of censoring, which occurs when the event of interest is not observed for some subjects during the study period [23]. Let $T \geq 0$ denote a non-negative random variable representing the survival time, the time until the occurrence of the event of interest. The SF $S(t)$ is defined as the probability that an individual survives beyond time t: $S(t) = \mathbb{P}(T > t)$. Due to censoring, the exact survival time may not be observed for a part of individuals. The most common form is right censoring, where the study ends before the event occurs or the individual is lost to follow-up. Let C denote the censoring time, and define the observed time $Y = \min(T, C)$ and the censoring indicator $\delta = \mathbb{I}\left(T \leq C\right)$, where $\delta = 1$ indicates that the event was observed and $\delta = 0$ indicates censoring.

When no covariates are present, the SF $S(t)$ is nonparametrically estimated by the Kaplan-Meier estimator [24]:

$$\hat{S}_{\mathrm{KM}}\left(t\right) = \prod_{j:t_{(j)} \leq t}\left(1 - \frac{d_j}{r_j}\right),$$

where d_j is the number of events at time $t_{(j)}$; r_j is the number of individuals at risk just before time $t_{(j)}$, i.e., the number of individuals with $Y > t_{(j)}$.

In most applications, interest lies not in the marginal survival distribution but in how survival depends on covariates $\boldsymbol{x} \in \mathbb{R}^p$. The conditional SF is defined as

$$S\left(t\middle|\boldsymbol{x}\right) = \mathbb{P}\left(T > t\middle|\boldsymbol{X} = \boldsymbol{x}\right).$$

Parametric approaches assume a specific functional form for $S(t \mid \boldsymbol{x})$, most notably the Cox proportional hazards model [6]. While the Cox model is widely used due to its interpretability and computational convenience, its proportional hazards assumption may be violated in practice. Nonparametric alternatives, such as the Beran estimator, offer greater flexibility by avoiding such parametric restrictions.

The Beran Estimator

The Beran estimator [5], also known as the conditional Kaplan-Meier estimator, provides a nonparametric estimate of the conditional SF $S(t \mid \boldsymbol{x})$ using kernel smoothing. It extends the Kaplan-Meier estimator to incorporate feature information by weighting observations according to their proximity in the feature space.

If our goal is to estimate SF $S(t \mid \boldsymbol{x}_0)$ for a feature vector $\boldsymbol{x}_0$, then observations with feature values close to $\boldsymbol{x}_0$ should receive greater weights than those far away, as they provide more relevant information about the conditional distribution. The Beran estimator allows us to extend the Kaplan-Meier estimator for accounting this idea.

Let $(\boldsymbol{x}_i, \delta_i, t_i)$, $i = 1, ..., n$, be instances from the dataset, where $t_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{I}\left(T_i \le C_i\right)$. We assume that $t_1 < t_2 < ... < t_n$ is an ordered sequence of times. Define a kernel function K_τ: $\mathbb{R}^p \to \mathbb{R}_+$ with bandwidth parameter $\tau > 0$ of the kernel. The kernel assigns higher weights to observations with $\boldsymbol{x}_i$ close to a target feature vector $\boldsymbol{x}$. The Beran estimator of the conditional SF $\hat{S}_n\left(t\middle|\boldsymbol{x}\right)$ is defined as:

$$\hat{S}_n\left(t\middle|\boldsymbol{x}\right) = \prod_{t_i \le t}\left\{1 - \frac{\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right)}{1 - \sum_{j=1}^{i-1}\alpha\left(\boldsymbol{x}, \boldsymbol{x}_j, \tau\right)}\right\}^{\delta_i},$$

where the weight $\alpha(\boldsymbol{x}, \boldsymbol{x}_i, \tau)$ can be defined through kernels as

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right) = \frac{K\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right)}{\sum_{j=1}^{n} K\left(\boldsymbol{x}, \boldsymbol{x}_j, \tau\right)}.$$

In particular, the "softmax" operation with a parameter τ is obtained when the Gaussian kernel is used, i.e., there holds

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right) = softmax\left(-\frac{\boldsymbol{x} - \boldsymbol{x}_i^2}{\tau}\right).$$

The practical performance of the Beran estimator critically depends on the choice of bandwidth τ. Cross-validation methods are commonly employed to optimize the model parameter τ.

While the Beran estimator with product kernel is widely used, its reliance on the Euclidean metric imposes several limitations. First, it assumes that the influence of each feature is separable and that features are independent in the distance calculation. As a results, the Euclidean metric ignores correlations among features, which can lead to suboptimal weighting in the presence of multicollinearity. Second, it is sensitive to differences in scale: features with larger magnitudes dominate the kernel weights unless carefully standardized. These limitations motivate the extension to the M-Beran estimator.

The M-Beran Estimator

The methodological core of our work is the incorporation of the Mahalanobis distance into the kernel weighting scheme of the Beran estimator. It is proposed to replace the Euclidean distance in the kernel functions of the Beran estimator with the Mahalanobis distance.

Let $\boldsymbol{M}$ be a positive-definite covariance matrix and $\tau > 0$ the bandwidth parameter. The M-Beran estimator of the conditional SF $S(t \mid \boldsymbol{x})$ is:

$$\hat{S}_n\left(t \mid \boldsymbol{x}\right) = \prod_{t_i \le t} \left\{ 1 - \frac{\alpha^M\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right)}{1 - \sum_{j=1}^{i-1} \alpha^M\left(\boldsymbol{x}, \boldsymbol{x}_j, \tau\right)} \right\}^{\delta_i},$$

where $\alpha^M(\boldsymbol{x}, \boldsymbol{x}_j, \tau)$ is a weight determined by the Mahalanobis distance from the target feature vector $\boldsymbol{x}$ to each observation's feature vector and determined for the Gaussian kernel as

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right) = softmax\left(-\frac{d_M^2\left(\boldsymbol{x}, \boldsymbol{x}_i\right)}{\tau} \right).$$

Given a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we first estimate the mean vector of features $\hat{\boldsymbol{\mu}}$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i.$$

Second, the sample covariance matrix $\boldsymbol{M}$ is estimated using the mean vector as:

$$\hat{\boldsymbol{M}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right)^T.$$

In high-dimensional settings where $p > n$, $\hat{\boldsymbol{M}}$ becomes singular, requiring regularization techniques. The next step is to compute the matrix $\boldsymbol{M}^{-1}$ by inverting the matrix $\hat{\boldsymbol{M}}$. The matrix $\hat{\boldsymbol{M}}$ in high-dimensional settings where $p > n$, may be not invertible. In this case, regularization can be used as follows:

$$\boldsymbol{M}_{reg} = \boldsymbol{M} + \varepsilon \boldsymbol{I},$$

were where ε is a small positive number, and $\boldsymbol{I}$ is the identity matrix.

Then we compute the Mahalanobis distance as:

$$d_M^2\left(\boldsymbol{x}, \boldsymbol{x}_i\right) = \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right)^T \boldsymbol{M}^{-1} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right).$$

The M-Beran estimator operates as follows: at each distinct event time t_j, we compute a kernel-weighted proportion of events, where the weights are determined by the Mahalanobis distance. Observations with $\boldsymbol{x}_i$ close to $\boldsymbol{x}$ in the Mahalanobis sense receive higher weight. The product of one minus these weighted proportions over all failure times up to t yields the estimated conditional survival probability.

The Mahalanobis distance generalizes the Euclidean distance, namely, if $\boldsymbol{M}$ is a diagonal matrix $\boldsymbol{M} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$, i.e., its diagonal consists of variances of individual features $\left(\boldsymbol{M} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)\right)$ and other elements are zero, then the Mahalanobis distance is equal to the standardized Euclidean distance.

Numerical results

To compare the M-Beran estimator with the standard Beran estimator, which uses the Euclidean distance in kernels, the following real benchmark datasets were used:

- Veterans' Administration Lung Cancer Study (*Veteran*) Dataset: 137 observations, 6 features;
- *Breast Cancer* Dataset: 198 observations, 80 features;
- *Lung Cancer* Dataset: 228 observations, 9 features;
- German Breast Cancer Study Group 2 (*GBSG2*) Dataset: 686 observations, 8 features;
- Monoclonal Gammopathy (*MGUS*) Dataset: 1384 observations, 5 features;
- Diabetic *Retinopathy* Dataset: 394 observations, 9 features;
- Ovarian Cancer Survival Data (*Ovarien*): 26 observations, 4 features.
- Chronic Granulomatous Disease (*CGD*) Dataset: 203 observations, 16 features.

In order to optimize the model temperature parameters τ of the estimators, we perform a 5-fold cross-validation on the training set. The cross-validation is performed with 100 repetitions. This procedure is realized by considering all possible values of τ in a predefined grid. The models are implemented by using the R software with the main R package "survival".

Table 1 presents a comparative analysis of the Mahalanobis-based Beran estimator (M-Beran) versus the standard Beran estimator across eight real survival datasets, using the C-index as the evaluation metric. The M-Beran estimator outperforms the standard Beran estimator on six out of eight datasets, demonstrating the general effectiveness of incorporating Mahalanobis distance into kernel-based survival estimation. The M-Beran achieves its highest C-index on the datasets which likely contain meaningful correlation structures among features that the Mahalanobis distance effectively captures. We can also note that the dataset *Breast Cancer* with the worse C-index of the M-Beran estimator has the large number of features (80 features) in comparison with other datasets, which may affect the estimation of the covariance matrix. The results indicate that the Mahalanobis distance is a valuable enhancement to the Beran estimator. The negative result on the *Breast Cancer* dataset serves as an important caveat, highlighting that the method is not universally superior and that the choice of distance metric should be guided by the underlying data structure.

Table 1

Results of comparison using the C-index metric for the M-Beran estimator with the standard Beran estimator on real datasets

Dataset	Standard Beran	M-Beran	Difference
Veteran	0.706	0.713	0.007
Breast Cancer	0.517	0.463	–0.054
GBSG2	0.513	0.523	0.01
Retinopathy	0.582	0.588	0.006
MGUS	0.724	0.742	0.018
Lung Cancer	0.646	0.668	0.022
Ovarien	0.774	0.798	0.024
CGD	0.583	0.724	0.141

In order to illustrate the difference in SFs predicted by the standard Beran estimator and the M-Beran estimator for the *Veteran* dataset, we show the corresponding SFs in Fig. 1, where SFs of the standard Beran estimator are depicted by solid lines, SFs of the M-Beran estimator are depicted by dashed lines. The left graph corresponds to the patients with: Karnofsky score – 90, Cell type – Squamous, Age – 55, Treatment – Standard, Prior therapy – No. The right graph corresponds to the patients with:

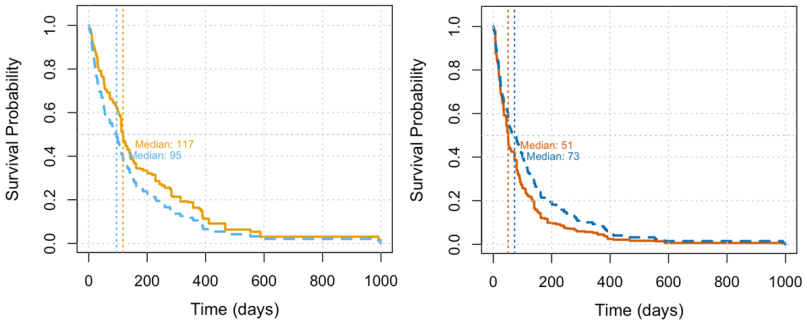


Fig. 1. SFs predicted by the standard Beran estimator and the M-Beran estimator for two patients from the Veteran dataset

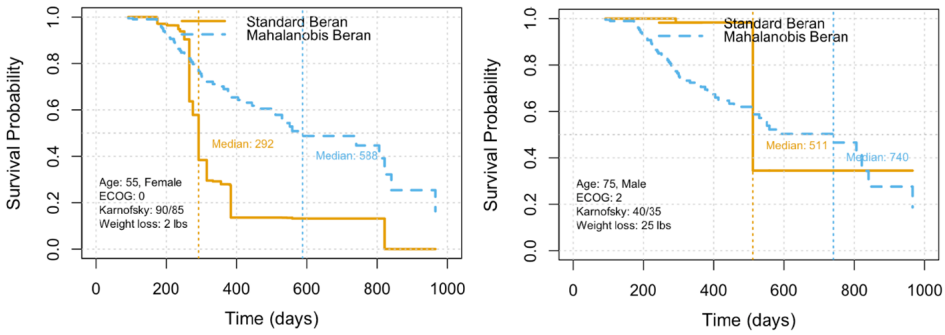


Fig. 2. SFs predicted by the standard Beran estimator and the M-Beran estimator for two patients from the Lung Cancer dataset

Karnofsky score – 40, Cell type – Adenocarcinoma, Age – 70, Treatment – Test, Prior therapy – Yes. It can be seen from Fig. 1 that the M-Beran estimator corrects SFs in accordance with the patient features.

Similar SFs of two patients from the *Lung Cancer* dataset are shown in Fig. 2. The left graph details a 55-year-old female with favorable health metrics, where the M-Beran estimator predicts a median survival of 588 days, nearly double the standard Beran model's 292 days. In this first scenario, the blue dashed line stays consistently higher than the orange line, indicating a superior predicted survival probability for the M-Beran approach. The right plot shows a 75-year-old male with poorer health indicators, where the M-Beran estimator again forecasts a longer median survival of 740 days compared to the standard model's 511 days. While the overall SF drops more rapidly for the second patient, the M-Beran estimator consistently estimates better outcomes than the standard Beran estimator in both distinct cases.

Fig. 3 illustrates the relationship between instances in the *Veteran* (the left graph) and *Lung Cancer* (the right graph) datasets by using the well-known procedure of the dimensionality reduction called t-SNE [25].

One can see from Fig. 3 that instances in the both datasets are highly correlated. This implies that the M-Beran estimator provides better results than the standard Beran estimator. The same can be seen from Table 1.

Fig. 4 shows the same relationship between instances in the *Breast Cancer* dataset. As a result, the standard Beran estimator outperforms the M-Beran estimator as it is also follows from Table 1.

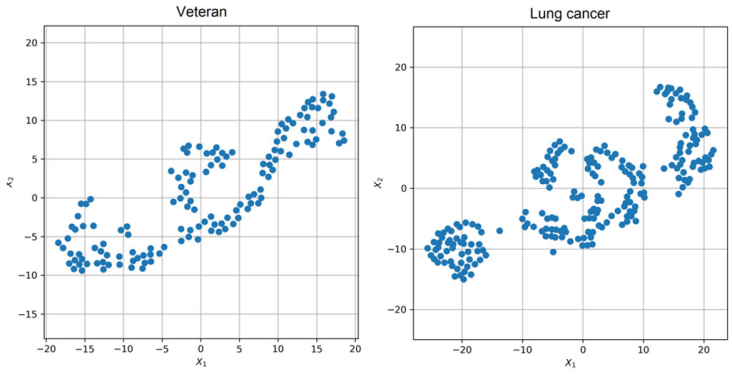


Fig. 3. T-SNE representation of the Veteran and Lung cancer datasets

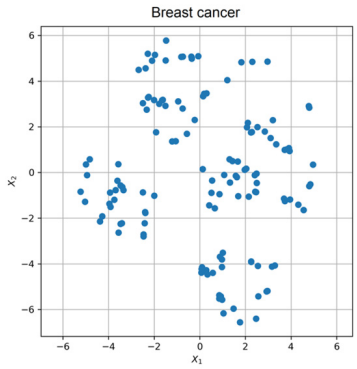


Fig. 4. T-SNE representation of the Breast Cancer dataset

It is interesting to note that the M-Beran estimator performance depends also on the kernel type. For example, let us consider the triangular kernel defined as follows:

$$K(u) = \max(0, 1 - |u|).$$

Hence, the weight α in the M-Beran estimator is

$$\alpha(\boldsymbol{x}, \boldsymbol{x}_i, \tau) = \frac{\max\left(0, 1 - d_M^2(\boldsymbol{x}, \boldsymbol{x}_i)/\tau\right)}{\sum_{j=1}^{n} \max\left(0, 1 - d_M^2(\boldsymbol{x}, \boldsymbol{x}_j)/\tau\right)}.$$

The use of this kernel significantly improves the M-Beran estimator performance for the *Veteran* dataset. The C-indices in this case for the standard Beran and M-Beran estimators are 0.582 and 0.724, respectively. At the same time, results for the *Lung Cancer* dataset (0.525 and 0.497) demonstrate deterioration in accuracy characteristics. This implies that the performance of the M-Beran estimator depends on many parameters, including the kernel type.

Discussion

Advantages

In the context of survival analysis, particularly when using nonparametric methods like the Beran estimator, the choice of distance metric is critical because it dictates how "similarity" between patients

or systems is defined. The Mahalanobis distance is important in survival analysis for the following key reasons:

1. Survival datasets often contain features measured in vastly different units (e.g., age in years, tumor size in millimeters, gene expression on a log scale, or sensor voltage). The Euclidean distance is sensitive to the scale of measurement. A variable with a large range will dominate the distance calculation over a variable with a small range, even if the latter is more predictive of survival. By incorporating the covariance matrix, it automatically standardizes the variables, ensuring that each covariate contributes appropriately to the distance metric regardless of its original unit or magnitude. This eliminates the need for arbitrary manual standardization or tuning separate bandwidths for each covariate.

2. In many survival applications (e.g., genomics, multi-sensor engineering monitoring), covariates are highly correlated. The Euclidean distance assumes orthogonality between dimensions. When two biomarkers are closely related, Euclidean distance counts the same information twice, making the similarity calculation inaccurate. The Mahalanobis distance accounts for the correlation structure of the data. It measures distance in terms of standard deviations from the mean, penalizing deviations that go against the natural correlation pattern of the data. This ensures that redundant information does not artificially inflate the similarity between observations.

3. Nonparametric survival estimators like the Beran estimator rely on kernel smoothing, where observations closer to the target vector $\boldsymbol{x}$ receive higher weights. Using Euclidean distance creates "spherical neighborhoods". In a correlated feature space, a spherical neighborhood may include observations that are statistically dissimilar while excluding those that are relevant but lie along a correlated axis. The Mahalanobis distance creates "elliptically symmetric neighborhoods" aligned with the data density. This ensures that the kernel weights are assigned to observations that are truly statistically similar, leading to a more accurate estimation of the conditional SF $S(t \mid \boldsymbol{X} = \boldsymbol{x})$.

4. Survival data is characterized by censoring, meaning information about the event time is missing for some subjects. This makes the effective sample size smaller and the estimation problem harder. Because censored data provides less information, it is crucial to maximize the utility of every uncensored neighbor. A superior distance metric ensures that the few available relevant neighbors are identified correctly. By reducing bias in neighbor selection, the Mahalanobis distance helps produce more stable survival curves even when data is sparse due to heavy censoring.

Challenges and disadvantages

While the Mahalanobis distance offers significant advantages over the Euclidean distance in handling scale and correlation, it also introduces specific challenges and disadvantages in the context of survival analysis:

1. The Mahalanobis distance requires the inversion of the covariance matrix $\boldsymbol{M}^{-1}$. In modern survival analysis, the number of features p often may exceed the number of observations n. When $p > n$, the covariance matrix is singular and cannot be inverted. Even when $p < n$, if features are nearly collinear, the matrix becomes ill-conditioned, leading to numerical instability and unreliable distance calculations. This requires regularization techniques which introduce additional tuning parameters.

2. The Mahalanobis distance relies on the sample mean and sample covariance matrix, both of which are highly sensitive to outliers. In survival data, extreme values can disproportionately skew the covariance estimate. This "masking effect" can cause the distance metric to shrink around the outliers, making actual outliers appear normal and distorting the kernel weights for the majority of observations. Robust estimators are required but are computationally more expensive.

3. The Mahalanobis distance typically uses a global covariance matrix estimated from the entire dataset. Survival relationships are often local; the correlation structure between features might differ across subgroups (e.g., the relationship between age and biomarkers might differ for early-stage versus late-stage disease). A global metric assumes the correlation structure is constant everywhere, which may

not align with the local nature of kernel-based estimators like the Beran estimator. A local covariance would be more appropriate but is difficult to estimate reliably due to data sparsity.

4. Computing the inverse covariance matrix has a computational complexity not less than $O(p^2)$, where p is the number of features. For datasets with a large number of features, this calculation becomes burdensome, especially if the distance needs to be recomputed repeatedly (e.g., during cross-validation). This is less of an issue with the Euclidean distance, which requires no matrix inversion.

5. Euclidean distance corresponds to intuitive geometric space (e.g., "5 years difference in age"). Mahalanobis distance measures distance in terms of standard deviations within a multivariate distribution. It is harder for clinicians or engineers to interpret what a specific Mahalanobis distance value means in practical terms. Explaining why two patients are "distant" requires explaining the underlying correlation structure, which can be a barrier to adoption in applied settings.

In sum, the Mahalanobis distance introduces computational burdens, instability in high dimensions, and sensitivity to outliers. In survival analysis, these limitations often require additional methodological steps, such as regularization, robust covariance estimation, or dimensionality reduction, to ensure the metric remains reliable.

REFERENCES

1. **Cox D.R.** *Analysis of survival data*. London: Chapman and Hall/CRC, 2018.

2. **Hosmer D.W., Lemeshow S., May S.** *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New Jersey: John Wiley & Sons, 2008. DOI: 10.1002/9780470258019

3. **Tattar P., Vaman H.J.** *Survival Analysis*. NY: Chapman and Hall/CRC, 2022. DOI: 10.1201/97810-03306979

4. **Wang P., Li Y., Reddy C.K.** Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys* (*CSUR*), 2019, Vol. 51, No. 6, Art. no. 110. DOI: 10.1145/3214306

5. **Beran R.** *Nonparametric regression with randomly censored survival data*, 1981.

6. **Cox D.R.** Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* (*Methodological*), 1972, Vol. 34, No. 2. Pp. 187–220. DOI: 10.1111/j.2517-6161.1972.tb00899.x

7. **Berrendero J.R., Bueno-Larraz B., Cuevas A.** On Mahalanobis Distance in Functional Settings. *Journal of Machine Learning Research*, 2020, Vol. 21, Pp. 1–33.

8. **Dabrowska D.M.** Non-Parametric Regression with Censored Survival Time Data. *Scandinavian Journal of Statistics*, 1987, Vol. 14, No. 3, Pp. 181–197.

9. **Dabrowska D.M.** Uniform Consistency of the Kernel Conditional Kaplan–Meier Estimate. *The Annals of Statistics*, 1989, Vol. 17, No. 3, Pp. 1157–1167.

10. **Ma Q., Zeng L., Tu S., Xu L.** Kernel Mean Matching with Mahalanobis Distance for Causal Inference of Time-to-event Outcome. *2022 IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), 2022, Pp. 509–514. DOI: 10.1109/BIBM55620.2022.9995021

11. **Baghfalaki T., Ganjali M., Berridge D.** Robust joint modeling of longitudinal measurements and time to event data using normal/independent distributions: a Bayesian approach. *Biometrical Journal*, 2013, Vol. 55, No. 6, Pp. 844–865. DOI: 10.1002/bimj.201200272

12. **Baghfalaki T., Ganjali M., Hashemi R.** Bayesian joint modeling of longitudinal measurements and time-to-event data using robust distributions. *Journal of Biopharmaceutical Statistics*, 2014, Vol. 24, No. 4, Pp. 834–855. DOI: 10.1080/10543406.2014.903657

13. **Liu M., Legault V., Fülöp T., Côté A.-M., Gravel D., Blanchet F.G., Leung D.L., Lee S.J., Nakazato Y., Cohen A.A.** Prediction of mortality in hemodialysis patients using moving multivariate distance. Frontiers in Physiology, 2021, Vol. 12, Art. no. 612494. DOI: 10.3389/fphys.2021.612494

14. **Roblin E., Cournède P.-H., Michiels S.** Evaluation of machine-learning models to measure individualized treatment effects from randomized clinical trial data with time-to-event outcomes. *arXiv:2506.12277*, 2025. DOI: 10.48550/arXiv.2506.12277

15. **Maas C.C.H.M., Kent D.M., Hughes M.C., Dekker R., Lingsma H.F., van Klaveren D.** Performance metrics for models designed to predict treatment effect. *BMC Medical Research Methodology*, 2023, Vol. 23, Art. no. 165. DOI: 10.1186/s12874-023-01974-w

16. **De Maesschalck R., Jouan-Rimbaud D., Massart D.L.** The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 2000, Vol. 50, No. 1, Pp. 1–18. DOI: 10.1016/S0169-7439(99)00047-7

17. **Anthony H., Kamnitsas K.** On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (*UNSURE 2023*), 2023, Vol. 14291, Pp. 136–146. DOI: 10.1007/978-3-031-44336-7_14

18. **Cabana E., Lillo R.E., Laniado H**. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, 2021, Vol. 62, Pp. 1583–1609. DOI: 10.1007/s00362-019-01148-1

19. **Colombo P., Dadalto E., Staerman G., Noiry N., Piantanida P.** Beyond Mahalanobis distance for textual OOD detection. *Advances in Neural Information Processing Systems*, 2022, Vol. 35, P. 17744–17759.

20. **Podolskiy A., Lipin D., Bout A., Artemova E., Piontkovskaya I.** Revisiting Mahalanobis distance for transformer-based out-of-domain detection. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021, Vol. 35, No. 35, Pp. 13675–13682.

21. **Brown P.O., Chiang M.C., Guo S., Jin Y., Leung C.K., Murray E.L., Pazdor A.G.M., Cuzzocrea A.** Mahalanobis distance based k-means clustering. *Big Data Analytics and Knowledge Discovery*, 2022, Vol. 13-428, Pp. 256–262. DOI: 10.1007/978-3-031-12670-3_23

22. **Nielsen S., Abdullaev L., Teo R., Nguyen T.M.** Elliptical attention. *38th Conference in Neural Information Processing Systems* (*NeurIPS*), 2024, Pp. 1–42.

23. **Wiegrebe S., Kopper P., Sonabend R., Bischl B., Bender A.** Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 2024, Vol. 57, Art. no. 65. DOI: 10.1007/s10462-023-10681-3

24. **Kaplan E.L., Meier P.** Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 1958, Vol. 53, No. 282, P. 457–481. DOI: 10.2307/2281868

25. **Van der Maaten L., Hinton G.** Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, Vol. 9, No. 86, P. 2579–2605.

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Lev V. Utkin
Уткин Лев Владимирович
E-mail: lev.utkin@gmail.com
ORCID: https://orcid.org/0000-0002-5637-1420

Andrei V. Konstantinov
Константинов Андрей Владимирович
E-mail: andrue.konst@gmail.com
ORCID: https://orcid.org/0000-0002-1542-6480

Stanislav R. Kirpichenko
Кирпиченко Станислав Романович
E-mail: kirpichenko.sr@gmail.com
ORCID: https://orcid.org/0000-0003-2275-1473

Natalia M. Verbova
Вербова Наталья Михайловна
E-mail: nag00@mail.ru
ORCID: https://orcid.org/0000-0001-8749-9470

Submitted: 07.04.2025; Approved: 17.06.2026; Accepted: 17.06.2026.
Поступила: 07.04.2025; Одобрена: 17.06.2026; Принята: 17.06.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 44–51.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 44–51.

Circuits and Systems for Receiving, Transmitting and Signal Processing

Устройства и системы передачи, приема и обработки сигналов

Short message
DOI: https://doi.org/10.18721/JCSTCS.19204
UDC 621.3.049.774.2

A FREQUENCY-SELECTIVE CIRCUIT IN 180 NM CMOS

M.M. Pilipko ✉ , *D.V. Morozov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ m_m_pilipko@rambler.ru

Abstract. A frequency-selective device has been developed in the form of a rejector active filter based on transconductance amplifiers and capacitors in the 180 nm CMOS technology of Mikron JSC. The layout dimensions are 62 x 26 μm. Post-layout simulation results are as follows: rejection frequency is 312.5 MHz; passband gain is 1.44 dB; rejection is 23.1 dB; power consumption is 4.8 mW; total harmonic distortion is 0.07%. A comparison with the results from other studies is performed. The device can be applied in optical communication systems.

Keywords: CMOS, frequency-selective device, rejector, active filter, transconductance amplifier, element imitation

Acknowledgements: The integrated circuit was manufactured using funds from the Ministry of Science and Higher Education of the Russian Federation within the framework of the federal project “Training of personnel and scientific foundation for the electronics industry” under the state assignment “Development of a prototyping methodology for electronic component base using domestic microelectronic manufacturing facilities based on MPW service”.

Citation: Pilipko M.M., Morozov D.V. A frequency-selective circuit in 180 nm CMOS. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 44–51. DOI: 10.18721/JCSTCS.19204

© Pilipko M.M., Morozov D.V., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Краткое сообщение
DOI: https://doi.org/10.18721/JCSTCS.19204
УДК 621.3.049.774.2

ЧАСТОТНО-ИЗБИРАТЕЛЬНОЕ УСТРОЙСТВО ПО ТЕХНОЛОГИИ КМОП 180 НМ

М.М. Пилипко ✉ , *Д.В. Морозов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ m_m_pilipko@rambler.ru

Аннотация. Разработано частотно-избирательное устройство в виде режекторного активного фильтра на основе транскондуктивных усилителей и конденсаторов в КМОП-технологии АО «Микрон» 180 нм. Размеры топологии 62 х 26 мкм. Результаты моделирования по топологии следующие: частота режекции 312,5 МГц; усиление в полосе пропускания 1,44 дБ; подавление 23,1 дБ; потребляемая мощность 4,8 мВт; коэффициент гармоник 0,07%. Проведено сравнение с результатами других работ. Устройство может быть применено в системах оптических коммуникаций.

Ключевые слова: КМОП, частотно-избирательное устройство, режектор, активный фильтр, транскондуктивный усилитель, элементная имитация

Финансирование: Производство интегральной микросхемы было выполнено за счет средств Министерства науки и высшего образования Российской Федерации в рамках федерального проекта «Подготовка кадров и научного фундамента для электронной промышленности» по государственному заданию «Разработка методики прототипирования электронной компонентной базы на отечественных микроэлектронных производствах на основе сервиса MPW».

Для цитирования: Pilipko M.M., Morozov D.V. A frequency-selective circuit in 180 nm CMOS // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 44–51. DOI: 10.18721/JCSTCS.19204

Introduction

Continuous-time Gm-C active filters based on transconductance amplifiers and capacitors represent frequency-selective circuits and are applied for high-frequency devices [1–7]. For the implementation of the Gm-C active filters, CMOS technology is usually used. In this paper, realization of a frequency-selective circuit as a rejector Gm-C active filter is discussed in the Mikron JSC 180 nm CMOS technology.

Transconductance amplifier circuit

The transconductance amplifier circuit is shown in Fig. 1. The transconductance amplifier is a voltage-controlled current source. It consists of an input stage [8] and a common-mode stabilization circuit. The input stage is based on two-finger MOSFETs M1–M6 and the common-mode circuit is realized by single-finger MOSFETs M7–M12. Inputs of the transconductance amplifier circuit are labeled “inp” – non-inverted and “inm” – inverted one. Outputs are “outp” – non-inverted and “outm” – inverted one. The common mode signal is “bias”. The transconductance amplifier circuit in the Gm-C active filter should have a tuning range to compensate temperature and CMOS technology variations in the production of the integrated circuit dies. The tuning capability of the transconductance amplifier circuit is provided by varying the voltage in the node “tune”. When using multiple transconductance amplifiers in an element imitator, the “com2” nodes of these amplifiers are connected together.

© Пилипко М.М., Морозов Д.В., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

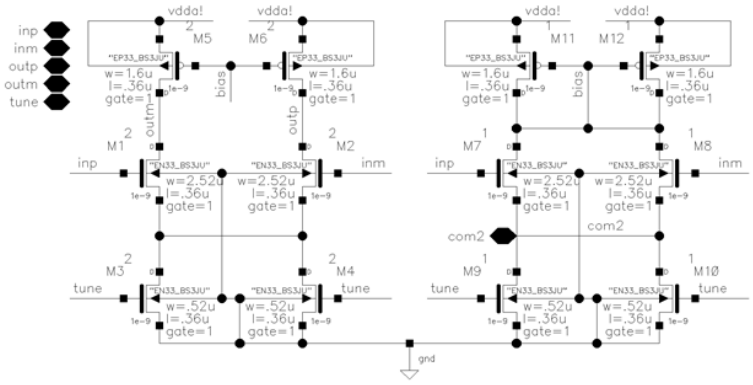


Fig. 1. Transconductance amplifier circuit

The transconductance amplifier circuit is implemented in the Mikron JSC 180 nm CMOS technology with a supply voltage of 3.3 V. Usage of a linear dropout regulator is implied with an output voltage of 2.8 V, which is indicated as “vdda!”. The channel dimensions of each MOSFET “w” and “l” are shown in Fig. 1. The transconductance of the amplifier circuit is near G_m = 242 uA/V at the nominal voltage value in the node “tune” equal to 1.4 V. The current consumption of the transconductance amplifier circuit is 230 uA.

Rejector Gm-C active filter

The second-order passive prototype of the rejector filter was used to implement the Gm-C active filter. The passive prototype is a balanced circuit, which is shown in Fig. 2 with a testbench and consists of input resistances R1 and R2, parallel oscillatory circuits L1C1 and L2C2, and a load output resistance R3. Inputs of the passive prototype are labeled “inp_RF” – non-inverted and “inm_RF” – inverted one. Outputs are “outp_RF” – non-inverted and “outm_RF” – inverted one. The common voltage, designated as “vcm”, is a half of the voltage in the node “vdda!”, i.e., 1.4 V. The rejection frequency is set as F_0 = 312.5 MHz. The element values are shown in fig. 2 and are as follows: R1 = = R2 = $1/(2G_m)$, R3 = $1/G_m$, C1/L1 = C2/L2 = G_m^2 and $F_0 = 1/(2\pi\sqrt{\text{L1C1}}) = 1/(2\pi\sqrt{\text{L2C2}})$. As the results of the AC-analysis, the amplitude and phase response of the passive prototype circuit are shown in Fig. 3. As can be seen, the passband attenuation of the passive prototype circuit is 6 dB.

The circuit of the rejector Gm-C active filter is shown in Fig. 4, *a*. It consists of seven transconductance amplifiers I1–I7 and four capacitors C1–C4. The input resistances R1 and R2 of the passive prototype circuit (Fig. 2) are implemented using transconductance amplifiers I1 and I2. The transconductance of the amplifier I1 is twice as much as G_m, which is indicated in Fig. 4, *a* as “2*Gm”, therefore, it is necessary to use channel widths “w” of all MOSFETs in the circuit of the amplifier I1 twice as much as in the circuit shown in Fig. 1. This increase in the transconductance of the amplifier I1 allows decreasing the input resistances and reducing the passband attenuation of the rejector Gm-C active filter.

The inductors L1 and L2 of the passive prototype circuit (Fig. 2) are implemented using two transconductance amplifiers and a capacitor for each inductor. Therefore, in Fig. 4, *a* transconductance amplifiers I3, I4 and the capacitor C3 are used as the inductor imitator of L1, while transconductance amplifiers I5, I6 and the capacitor C4 are used as the inductor imitator of L2. Considering C3 = L1G_m^2 and C4 = L2G_m^2, at C1/L1 = C2/L2 = G_m^2 all capacitors have the same capacitance, i.e., C1 = C2 = C3 = C4 =123 fF.

The load output resistance R3 of the passive prototype circuit (Fig. 2) is implemented using the transconductance amplifier I7 in Fig. 4, *a*. The transconductance of the amplifier I7 is half as much as

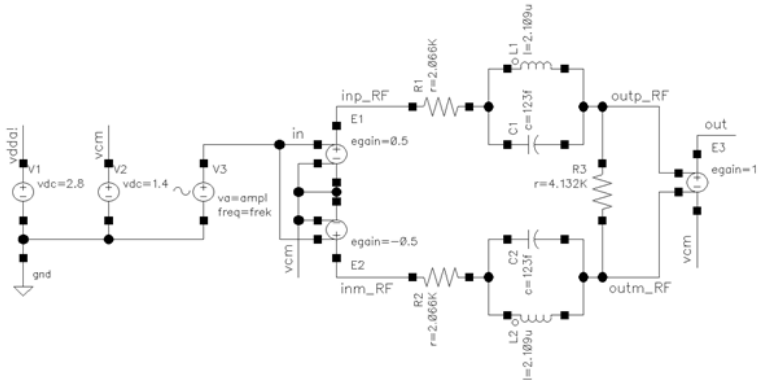


Fig. 2. Testbench with the second-order passive prototype of the rejector filter

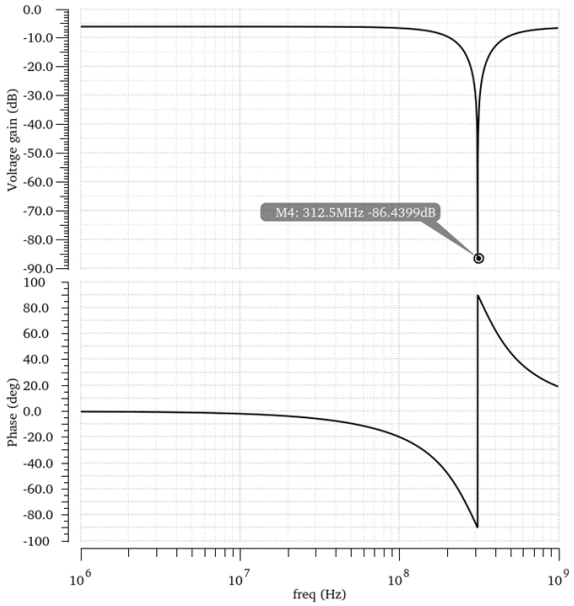


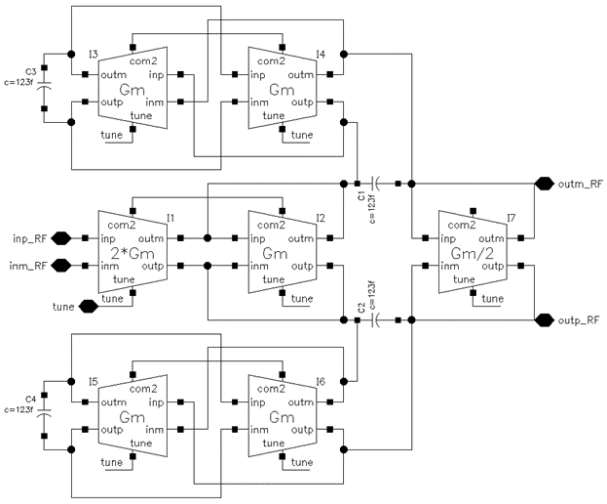
Fig. 3. Amplitude and phase response of the passive prototype circuit

G_m, which is indicated in fig. 4a as "Gm/2", therefore, it is necessary to use single-finger MOSFETs M1–M6 in the input stage of the amplifier I7 in comparison with the circuit shown in Fig. 1. This decrease in the transconductance of the amplifier I7 allows increasing the load output resistance and reducing passband attenuation of the rejector Gm-C active filter.

The layout of the rejector Gm-C active filter is shown in Fig. 4, *b*. Sizes of the layout are 62 x 26 μm. As the results of the AC-analysis, the amplitude and phase response of the post-layout simulation are shown in Fig. 5. As can be seen, the passband gain of the circuit after extraction is 1.44 dB.

The amplitude response shown in Fig. 5 has been acquired at the "typ" corner, i.e., nominal values of MOSFETs and capacitors, with the tuning voltage V_{tune} = 1.4 V. In the case of the "slow3s" corner at V_{tune} = 1.4 V, the rejection frequency has changed to F_0 = 264 MHz. To have the nominal value of the rejection frequency, it is necessary to change the tuning voltage to V_{tune} = 1.65 V. In the case of the "fast3s" corner at V_{tune} = 1.4 V, the rejection frequency has changed to F_0 = 403 MHz. To have the nominal value of the rejection frequency, it is necessary to change the tuning voltage to V_{tune} = = 1.21 V. As can be seen, a feasible change of the tuning voltage allows to adjust the filter circuit to the nominal rejection frequency.

a)



b)

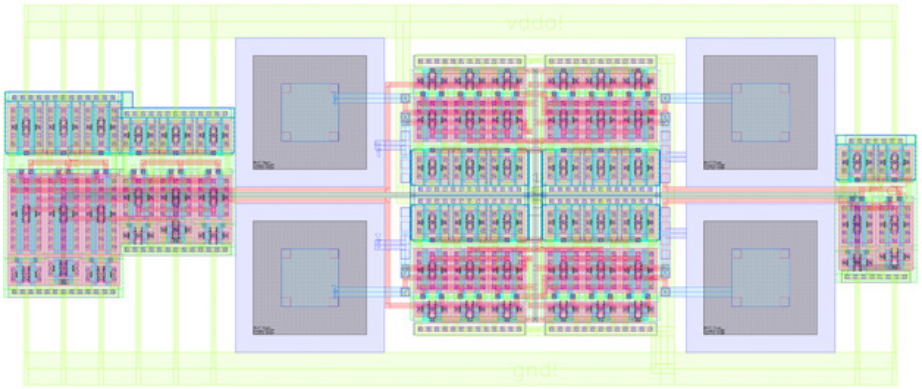


Fig. 4. Rejector Gm-C active filter

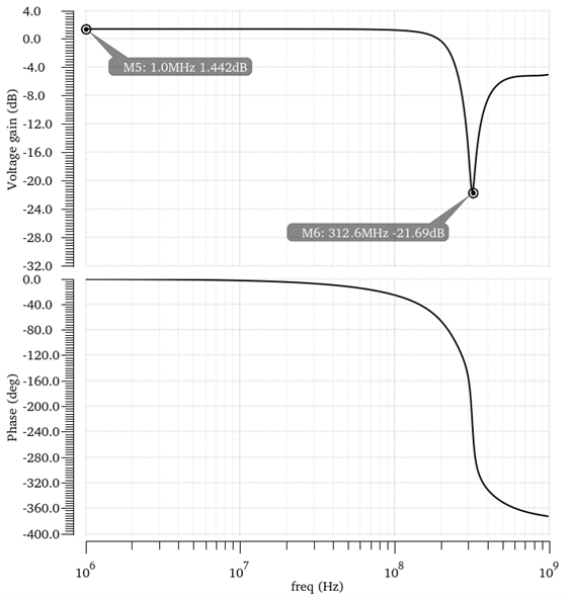


Fig. 5. Amplitude and phase response of the post-layout simulation

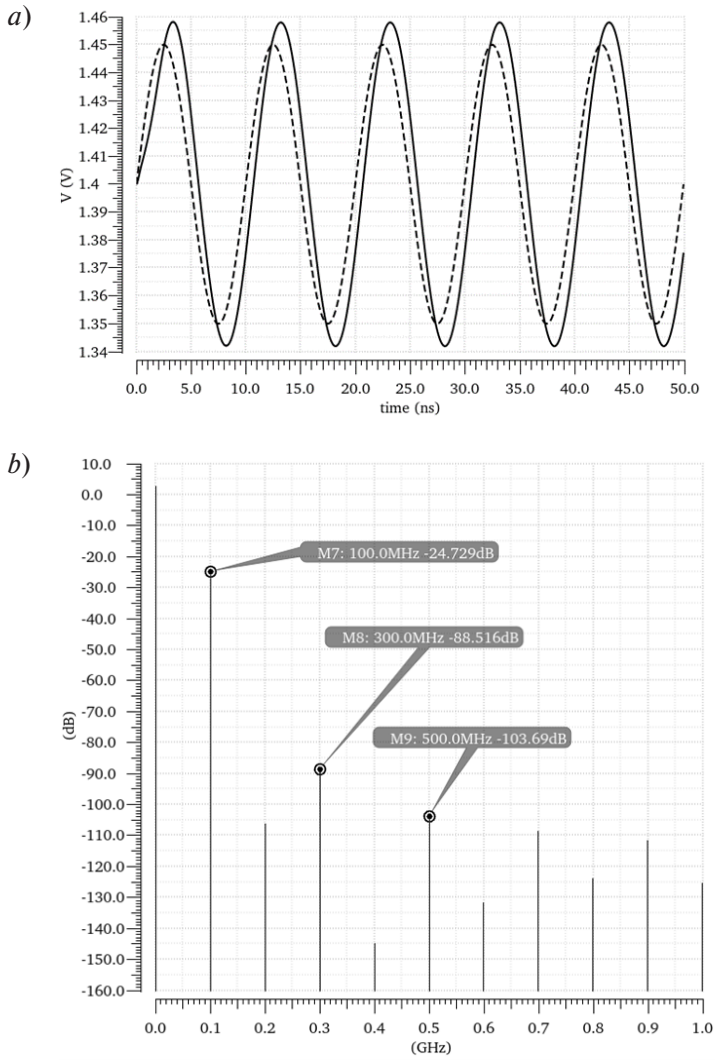


Fig. 6. Differential input and output signals of the post-layout simulation and the output signal spectrum

As the results of the transient analysis, at the frequency of 100 MHz the differential input (the dash line) and output (the solid line) signals of the post-layout simulation and the output signal spectrum are shown in Fig. 6. As can be seen, the nonlinear distortion of the layout is no more than 63 dB, which corresponds to the value of total harmonic distortion equal to 0.07%. Main characteristics of the realized rejector Gm-C active filter are summarized in table 1 in comparison with the results published in [3].

Table 1

Comparison of post-layout simulation results

Main characteristics	This work	[3]
CMOS technology	180 nm	130 nm
Layout area	0.002 mm^2	0.046 mm^2
Number of transconductance amplifiers in the filter circuit	7	6
Power consumption	4.8 mW	3.5 mW

End of Table 1

Main characteristics	This work	[3]
Power consumption per one transconductance amplifier	0.7 mW	0.6 mW
3-dB bandwidth	225 MHz	156 MHz
Attenuation	23.1 dB	28.9 dB
In-band ripple	0 dB	1 dB
Total harmonic distortion	0.07%	0.5%

Conclusion

The realization of a rejector Gm-C active filter in the Mikron JSC 180 nm CMOS technology has been discussed. In comparison with the results published in [3], this work provides a more compact layout, a higher frequency range, and a lower total harmonic distortion at a similar power consumption per one transconductance amplifier. The circuit is applicable in networking and optical communication.

REFERENCES

1. **Pérez-Bailón J., Calvo B., Medrano N., Ramírez-Angulo J.** A 1.8 V Gm-C Highly Tunable Low Pass Filter for Sensing Applications. *2020 IEEE International Instrumentation and Measurement Technology Conference* (*I2MTC*), 2020, Pp. 1–5. DOI: 10.1109/I2MTC43012.2020.9128692

2. **Ghasemi A.R., Aminzadeh H., Ballo A.** Ladder-Type Gm-C Filters with Improved Linearity. *IEEE Access*, 2023, Vol. 11, Pp. 41503–41513. DOI: 10.1109/ACCESS.2023.3270372

3. **Wang J.-f., Liu X.-f., Ju S.-r., Wang T., Qian J.** Design of a low-pass filter for wideband transceiver. *2023 8th International Conference on Integrated Circuits and Microsystems* (*ICICM*), 2023, Pp. 397–400. DOI: 10.1109/ICICM59499.2023.10366038

4. **Corbacho I., Carrillo J.M., Ausín J.L., Domínguez M.Á., Peréz-Aloe R., Duque-Carrillo J.F.** A 4-Decade-Tunable High-Selectivity G_m-C Bandpass Filter for Simultaneous Multi-Sine Bioimpedance Analysis. *IEEE Transactions on Circuits and Systems – I: Regular Papers*, 2024, Vol. 71, No. 1, Pp. 160–173. DOI: 10.1109/TCSI.2023.3331979

5. **Corbacho I., Carrillo J.M., Ausín J.L., Domínguez M.Á., Duque-Carrillo J.F.** 0.8-V CMOS G_m-C Bandpass Filter for Electrical Bioimpedance Spectroscopy. *2021 28th IEEE International Conference on Electronics, Circuits, and Systems* (*ICECS*), 2021, Pp. 1–4. DOI: 10.1109/ICECS53924.2021.9665648

6. **Corbacho I., Carrillo J.M., Ausín J.L., Domínguez M.A., Pérez-Aloe R., Duque-Carrillo J.F.** Impact of Process Variations on the Performance of a Widely Tunable CMOS Transconductor. *2022 18th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design* (*SMACD*), 2022, Pp. 1–4. DOI: 10.1109/SMACD55068.2022.9816190

7. **Corbacho I., Carrillo J.M., Ausín J.L., Domínguez M.Á., Pérez-Aloe R., Duque-Carrillo J.F.** Low-Gm CMOS Transconductors with Wide Tuning Range for Bioimpedance Spectroscopy. *2023 19th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design* (*SMACD*), 2023, Pp. 1–4. DOI: 10.1109/SMACD58065.2023.10192142

8. **Morozov D.V., Korotkov A.S.** A realization of low-distortion CMOS transconductance amplifier. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 2001, Vol. 48, No. 9, Pp. 1138–1141. DOI: 10.1109/81.948443

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Mikhail M. Pilipko
Пилипко Михаил Михайлович
E-mail: m_m_pilipko@rambler.ru
ORCID: https://orcid.org/0000-0003-3813-6846

Dmitry V. Morozov
Морозов Дмитрий Валерьевич
E-mail: dvmorozov@inbox.ru
ORCID: https://orcid.org/0000-0003-3403-0120

Submitted: 06.03.2026; Approved: 28.04.2026; Accepted: 19.05.2026.
Поступила: 06.03.2026; Одобрена: 28.04.2026; Принята: 19.05.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 52–60.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 52–60.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19205
UDC 621.372.552

DESIGN OF POWER LIMITER FOR ANTENNA-FEEDER SYSTEM FOR DIRECTION FINDING PROTECTION ON GaAs pHEMT

D.V. Klimenko[1], *A.B. Nikitin*[2], *A.A. Stroganov*[2] ✉

[1] Special Technology Center (STC LLC), St. Petersburg, Russian Federation;
[2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

✉ stroganov187@yandex.ru

Abstract. This paper presents the development results of a power limiter for the 30–6500 MHz frequency range, implemented as a microwave monolithic integrated circuit (MMIC) based on a domestic GaAs pHEMT technology. Implementing the power limiter as a monolithic integrated circuit reduces the weight and size of the final device compared to a hybrid implementation. The power limiter consists of three cascaded stages connected in parallel, comprising Schottky diodes interconnected by microstrip transmission lines that compensate for the parasitic capacitances of the diodes. The developed power limiter exhibits an insertion loss of less than 0.7 dB, an input and output VSWR of less than 1.2, an input 1 dB compression point of at least 22.8 dBm, and a maximum input power of at least 40.7 W. The MMIC die dimensions are 1400 × 1300 μm. The developed power limiter can be used in the input stage of receiving devices, such as antenna feeder systems for direction finding.

Keywords: microwave, monolithic integrated circuit, power limiter, antenna-feeder system, GaAs pHEMT, Schottky diode

Acknowledgements: The integrated circuit was manufactured using funds from the Ministry of Science and Higher Education of the Russian Federation within the framework of the federal project “Training of personnel and scientific foundation for the electronics industry” under the state assignment “Development of a prototyping methodology for electronic component base using domestic microelectronic manufacturing facilities based on MPW service” (FSMR-2023-0008).

Citation: Klimenko D.V., Nikitin A.B., Stroganov A.A. Design of power limiter for antenna-feeder system for direction finding protection on GaAs pHEMT. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 52–60. DOI: 10.18721/JCSTCS.19205

© Klimenko D.V., Nikitin A.B., Stroganov A.A., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19205
УДК 621.372.552

РАЗРАБОТКА ОГРАНИЧИТЕЛЯ МОЩНОСТИ ДЛЯ ЗАЩИТЫ АНТЕННО-ФИДЕРНЫХ СИСТЕМ ПЕЛЕНГОВАНИЯ НА ОСНОВЕ GaAs pHEMT-ТЕХНОЛОГИИ

Д.В. Клименко[1], А.Б. Никитин[2], А.А. Строганов[2] ✉

[1] Специальный Технологический Центр (ООО «СТЦ»), Санкт-Петербург, Российская Федерация;

[2] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

✉ stroganov187@yandex.ru

Аннотация. В данной статье представлены результаты разработки ограничителя мощности диапазона частот 30–6500 МГц, реализованного в виде СВЧ монолитной интегральной схемы на основе отечественной GaAs pHEMT-технологии. Реализация ограничителя мощности в виде монолитной интегральной схемы позволяет уменьшить массогабаритные параметры конечного изделия по сравнению с гибридным вариантом реализации. Ограничитель мощности состоит из трех каскадов, включенных параллельно, диодов Шоттки, соединенных микрополосковыми линиями передач, компенсирующими влияние паразитных емкостей диодов. Разработанный ограничитель мощности обладает вносимыми потерями менее 0,7 дБ, КСВН входа и выхода составляет менее 1,2, величина входной мощности точки компрессии по уровню 1 дБ составляет не менее 22,8 дБм, максимальная входная мощность не менее 40,7 Вт. Размеры кристалла монолитной интегральной схемы составили 1400 х 1300 мкм. Разработанный ограничитель мощности может быть использован во входном тракте приемных устройств, таких как антенно-фидерные системы пеленгования.

Ключевые слова: СВЧ, монолитная интегральная схема, ограничитель мощности, антенно-фидерная система, GaAs pHEMT, диод Шоттки

Финансирование: Производство интегральной микросхемы было выполнено за счет средств Министерства науки и высшего образования Российской Федерации в рамках федерального проекта «Подготовка кадров и научного фундамента для электронной промышленности» по государственному заданию «Разработка методики прототипирования электронной компонентной базы на отечественных микроэлектронных производствах на основе сервиса MPW» (FSMR-2023-0008).

Для цитирования: Klimenko D.V., Nikitin A.B., Stroganov A.A. Design of power limiter for antenna-feeder system for direction finding protection on GaAs pHEMT // Computing, Telecommunications and Control. 2026. T. 19, № 2. C. 52–60. DOI: 10.18721/JCSTCS.19205

Introduction

In modern antenna-feeder bearing systems (AFS) for direction finding, the power limiter (PL) is one of the key components of the input stage [1]. The primary threat to AFS is a strong signal that falls within the system's operating frequency band. This could be either random interference from a nearby transmitter or a deliberate attempt to "blind" the bearing system through a targeted attack. Such a powerful signal can overload the receiving stage elements (for example, a low-noise amplifier in a preselector) or cause irreversible damage to the device (Fig. 1). Consequently, the PL's role is to safeguard highly sensitive components against signals whose power surpasses a safe threshold [1, 2].

One of the major trends in the evolution of radio electronic equipment is the adoption of microwave monolithic integrated circuits (MMICs) [3, 4]. Specifically, PLs implemented using MMICs

© Клименко Д.В., Никитин А.Б., Строганов А.А., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

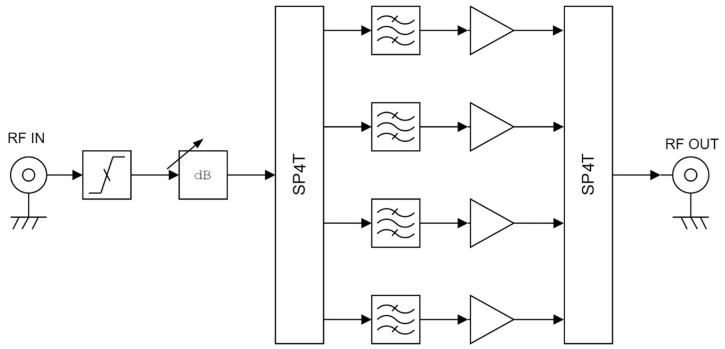


Fig. 1. Preselector block diagram

are characterized by compact dimensions, high performance, and the absence of the need for external matching [5, 6].

Due to difficulties in accessing high-tech foreign-produced equipment and the limited availability of specialized microcircuit solutions, there is a particular focus on expanding the range of MMICs manufactured using domestic technological processes (for example, at Svetlana-Rost JSC) [7–10].

This work explores the specifics of designing a power limiter for the frequency range from 30 to 6500 MHz based on domestic GaAs pHEMT technology, utilizing process design kits for solid-state microwave devices (PDK_pHEMT05D) [11].

The design of a broadband MMIC PL based on domestic GaAs pHEMT technology presents several challenges. The primary scientific and technical problem lies in the contradiction between the need for high input power handling (tens of watts) and the low breakdown voltage of Schottky diodes compared to PIN-diodes. Furthermore, ensuring minimal insertion loss (< 1 dB) and voltage standing wave ratio (VSWR) < 1.5 across a wide frequency range (30–6500 MHz) is non-trivial due to the parasitic capacitance of the diodes.

The purpose of this work is to develop a compact MMIC PL for antenna-feeder bearing systems using domestic GaAs pHEMT technology, providing protection against high-power input signals while maintaining linearity in the low-signal mode.

To achieve this goal, the following research tasks were formulated:

1. To select and justify the choice of the domestic technological process (pHEMT05 vs pHEMT025) based on the analysis of the current-voltage characteristics of Schottky diodes.

2. To determine the optimal number of limiting stages and the periphery of the diodes to ensure the required compression point (P1dB ≥ 20 dBm) and maximum input power (≥ 10 W).

3. To develop a distributed structure of the limiter where the parasitic capacitance of the diodes is absorbed by the microstrip transmission lines to achieve broadband matching.

4. To evaluate the key electrical parameters (insertion loss, VSWR, P1dB) through electromagnetic simulation and compare them with existing commercial and domestic analogs.

Operating principle

The PL MMIC is a device composed of several stages implemented as segments of a transmission line with nonlinear elements – diodes – that alter their electrical characteristics when the incident microwave power exceeds a certain threshold level [1, 3, 5, 7]. In the high-power mode, when the input microwave power equals or surpasses the threshold level, the PL reduces it to a level safe for the receiver's input circuits. Due to its relatively low value, this power is referred to as the leakage power. In the low-power mode, the input microwave power is significantly lower than the threshold level, and the device allows the input signal to pass through with minimal attenuation [1, 13].

For PL MMIC, PIN-diodes or Schottky diodes are utilized [5, 6]. Schottky diodes have a lower breakdown voltage and are less resistant to overheating compared to PIN-diodes [12]. However, they are technologically compatible with low-noise transistors, unlike PIN-diodes, enabling the creation of core-chips [4, 6] and reducing the cost of the final product.

PL design

At present, Svetlana-Rost JSC offers two GaAs technological processes with diodes for the commercial launch of MMICs: pHEMT05 (0.5 μm gate length) and pHEMT025 (0.25 μm gate length).

Analysis of the diode IV-curves (Fig. 2) revealed that the breakdown voltage for the first process is approximately -11.2 V, whereas for the second process it is around -5.7 V. Consequently, pHEMT05 is a more preferable choice, as diodes produced using this process feature significantly higher breakdown voltage and can thus handle greater input microwave power.

In addition to the breakdown voltage, the performance of the PL is influenced by the number of diodes in the stage and the periphery of the diodes used [6, 13]. A larger diode periphery enables higher microwave power handling. However, the higher parasitic capacitance of such diodes restricts the operating frequency range. To address this issue, a distributed structure was employed, where the parasitic capacitance of the diodes is absorbed by a distributed transmission line [14, 15].

The methodology for determining the number of stages and the total diode periphery was based on the required input 1 dB compression point (P1dB) and the withstand input power. A single-stage limiter using Schottky diodes with a total periphery of 640 μm (4 x 160 μm) provides a P1dB of approximately 26–26.5 dBm but fails to limit power above 30 dBm due to the risk of breakdown. To achieve withstand input power no less than 10 W, a three-stage distributed topology was adopted (Fig. 4). This allows for sequential activation of the diode cells: diodes VD11–VD26 are turned on first and handles moderate over loads, while diodes VD1–VD10 provide limiting at high power levels, preventing voltage swings from exceeding the reverse breakdown voltage of 11.2 V (as verified in Fig. 9).

Microstrip transmission lines composed of two metallization layers are used as inductances L1 and L2. The characteristic impedance and electrical length of each transmission line section were synthesized to absorb the parasitic capacitance Cj into a distributed LC network, forming a low-pass filter structure. The optimization goal was to achieve insertion loss < 1 dB and VSWR < 1.5 over the 30–6500 MHz band. The two-layer metal lines were chosen not only for high current density but also to reduce the series inductance per unit length, which improves the cutoff frequency of the limiter.

The layout of the designed PL is presented in Fig. 4. The MMIC dimensions are 1.4 x 1.3 mm.

All electromagnetic simulations of the passive structures and harmonic balance simulations of the nonlinear circuit (Schottky diodes) were performed using AWR Design Environment. The performance of the PL in the small-signal mode is demonstrated in Figs. 5–6. Insertion loss is no more than

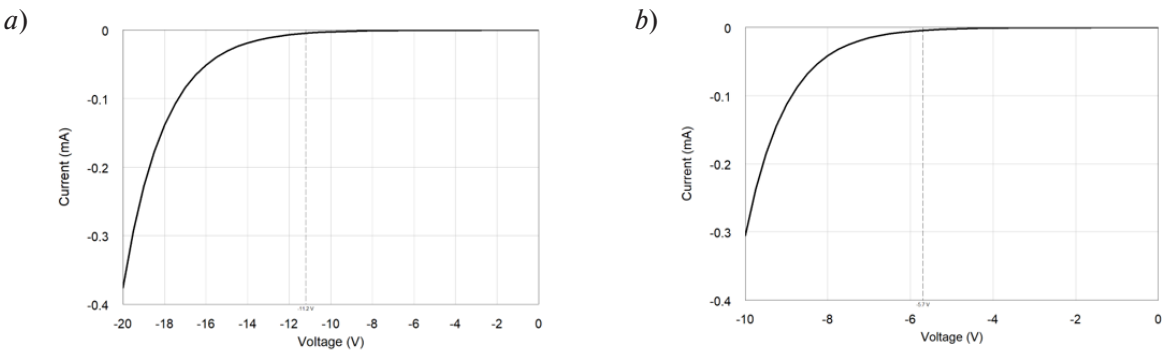


Fig. 2. IV-Curve of the 4 x 40 μm reverse biased diode: pHEMT05 (*a*); pHEMT025 (*b*)

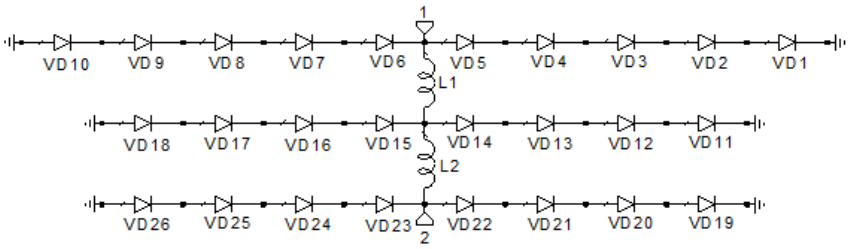


Fig. 3. PL schematic

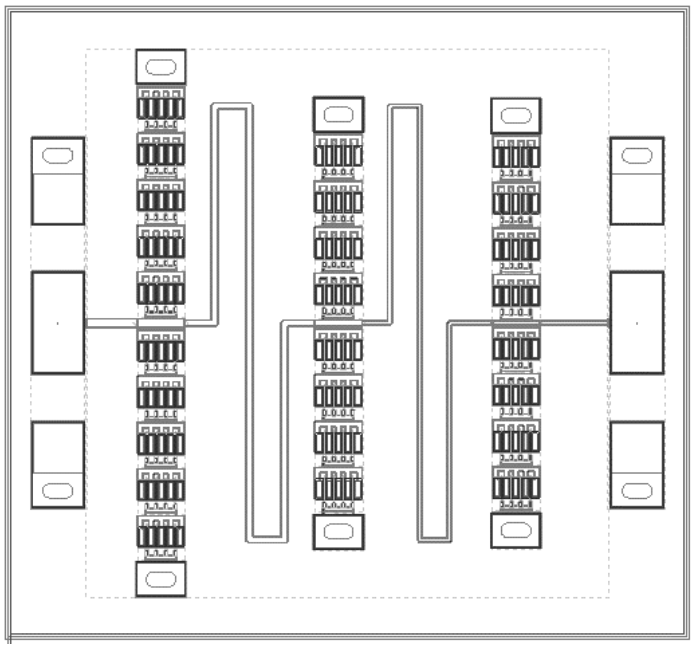


Fig. 4. PL layout

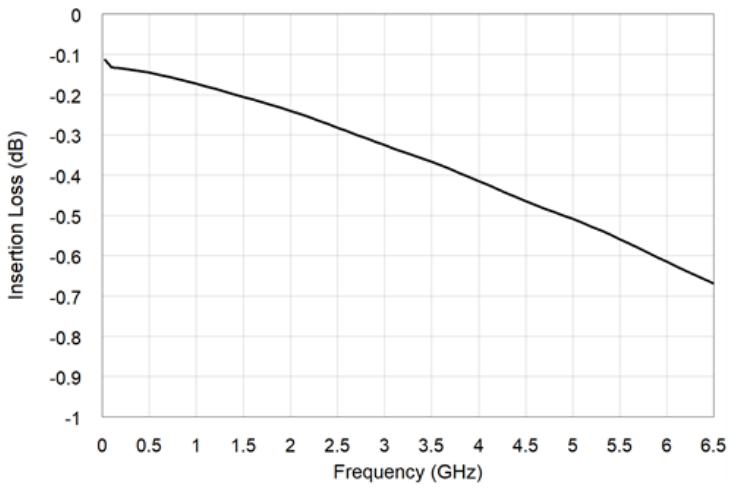


Fig. 5. PL insertion loss

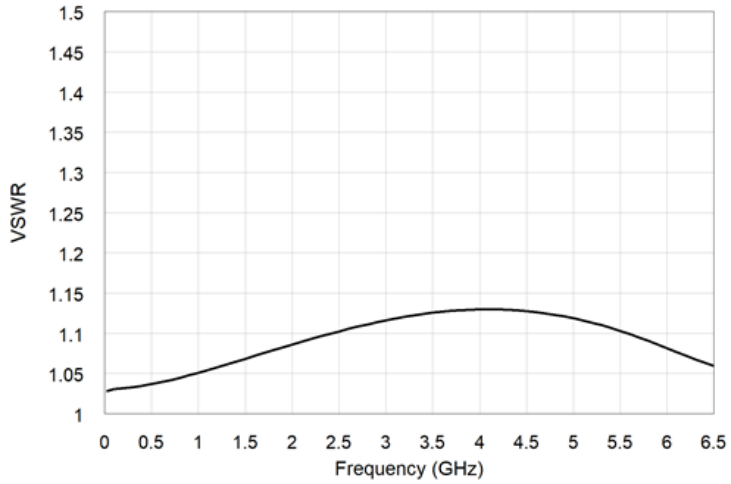


Fig. 6. PL VSWR

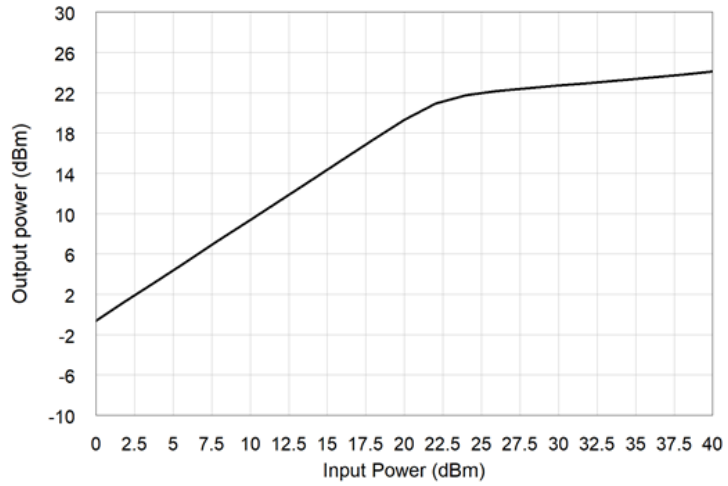


Fig. 7. Output power vs input power at 6 GHz

0.7 dB, and VSWR does not exceed 1.2 units. These parameters consider the parasitic impact of wire bonding using 35 μm diameter gold wire.

The results of the analysis of the PL in the high-signal mode are shown in Figs. 7–8.

Fig. 9 shows the simulated voltage swings across the VD5 and VD6 diodes when the input power is 10 W (40 dBm) at 6 GHz. As you can see from the graph, the peak voltage is almost 8 V, which is lower than the breakdown voltage of 11.2 V. The predicted power of this limiter is 40.7 W (46.1 dBm).

A comparison of the results shows that the main characteristics of the PL designed on the basis of domestic technology are comparable to foreign and domestic analogues, and in some cases even surpasses them. For example, the Qorvo TGL2210-SM[1] GaAs PL is also based on three cascades. With a similar operating frequency range (0.05–6.0 GHz) and comparable insertion loss (less than 0.7 dB), this PL provides a higher maximum input power (49.3 dBm) due to the use of PIN-diodes. However, Qorvo's chip has significantly larger dimensions (4.0 x 4.0 mm), and the linearity level (input 1 dB compression point is less than 20 dBm) does not allow it to be used in AFS due to possible input signal distortions.

[1] TGL2210-SM, Available: https://www.qorvo.com/products/p/TGL2210-SM (Accessed 17.02.2026)

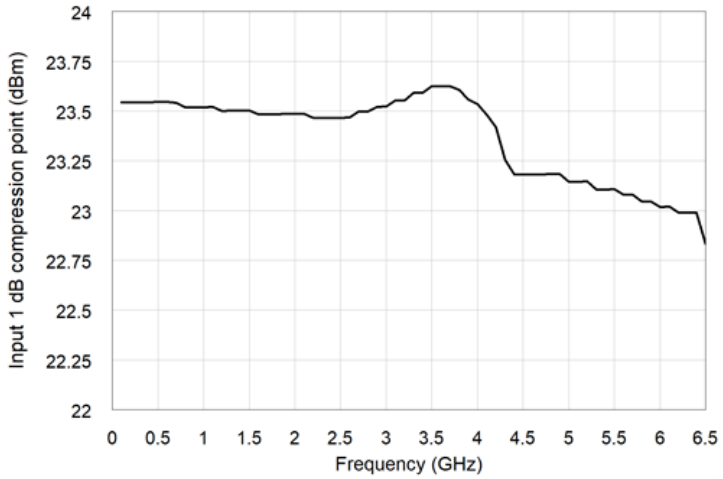


Fig. 8. Input 1 dB compression point vs frequency

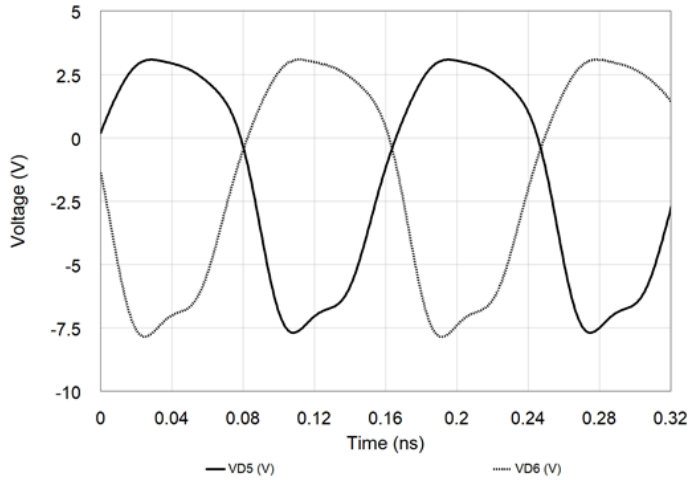


Fig. 9. Voltage swings on VD5 and VD6 diodes at 40 dBm input power

The domestic analog, the M54404 PL from Planet Argal[2], is also comparable in the operating frequency range (0.1–6.0 GHz), but in other characteristics it is completely inferior to the developed PL: insertion loss reaches 1.5 dB, input 1 dB compression point is less than 20 dBm, maximum input power is not more than 40 dBm, and the dimensions of the MMIC are 3.4 x 2.3 mm.

Conclusion

As a result of this research, the PL MMIC was designed on the basis of domestic GaAs pHEMT-technology using process design kit pHEMT025D. The PL consists of three stages of Schottky diodes connected in parallel by microstrip transmission lines. The values of the parameters of the components have been found that provide insertion losses of less than 0.7 dB in the frequency range 30–6500 MHz, VSWR of no more than 1.2, input 1 dB compression point is no less than 22.8 dBm, maximum input power of at least 40.7 W. The size of the MMIC is 1400 x 1300 microns.

[2] SVCh zashchitnoe ustroistvo M54404 [Microwave protective device M54404], Available: https://argall.nov.ru/wp-content/uploads/2021/08/M54404.pdf (Accessed 17.02.2026)

The designed PL is comparable in basic parameters to its analogs, but is significantly smaller in size and has a higher level of linearity. This PL is planned to be used in the AFS preselector of domestic production.

REFERENCES

1. **Cherepanov V.P., Posysaev E.I.** *Zashchita radioelektronnoy apparatury ot elektricheskikh peregruzok* [*Protection of electronic equipment from electrical overloads*]. Moscow: IP RadioSoft, 2010.

2. **Kantiuk D.V., Volnianskii N.O., Nikul'chenko N.N., Tolstolutskaia A.V., Tolstolutskii S.I.** Optimizatsiia tekhnologii izgotovleniia MIS SVCh ogranichitelei moshchnosti na arsenide galliia dlia modulei AFAR [Optimization of manufacturing technology for MMIC power limiters based on gallium arsenide for APAR modules]. *Sbornik trudov I Vserossiiskoi nauchno-tekhnicheskoi konferentsii "Electronika i mikroelectronika SVCh"* [*Proceedings of the I All-Russian Scientific and Technical Conference "Microwave Electronics and Microelectronics"*], 2012, Pp. 97–101.

3. **Belous A.I., Merdanov M.K., Shvedov S.V.** *SVCh-elektronika v sistemah radiolokatsii I sviazi* [*Microwave electronics in radar and communication systems*]. Moscow: Tekhnosfera, 2018.

4. **Vikulov I.** Monolithic microwave ICs the technology basis of AESA. *Electronics: Science, Technology, Business*, 2012, Vol. 17, No. 7, Pp. 60–73.

5. **Yuschenko A.Yu., Ayzenshtat G.I., Monastyrev E.A., Bozhkov V.G.** MMIC's broadband switches and limiters with heterostructured GaAs pin-diodes. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki*, 2011, Vol. 24, No. 2, Pp. 10–16.

6. **Dobush I.M.** Review of designing principles and circuitry of MMIC limiters. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki*, 2010, Vol. 22, No. 2, Pp. 44–48.

7. **Krasovitskii D. M., Filaretov A.G., Chaly V.P.** Pervaia otechestvennaia SVCH-fabrika – "faundri" s voennoi priemkoi: metodologiia, problemy i perspektivy [The first domestic microwave factory – "foundry" with military acceptance: methodology, problems and prospects]. *SVCh-elektronika* [*Microwave Electronics*], 2020, Vol. 5, No. 1, Pp. 16–19.

8. **Klimenko D.V., Nikitin A.B., Stroganov A.A., Tsikin I.A.** Design of monolithic microwave integrated circuit bandpass filters on GaAs pHEMT. *2023 International Conference on Electrical Engineering and Photonics* (*EExPolytech*), 2023, Pp. 10–13. DOI: 10.1109/EExPolytech58658.2023.10318671

9. **Klimenko D.V., Nikitin A.B., Stroganov A.A., Tsikin I.A.** Features of centimeter-band filter-bank design based on GaAs pHEMT-technology. *Computing, Telecommunications and Control*, 2023, Vol. 16, No. 3, Pp. 7–17. DOI: 10.18721/JCSTCS.16301

10. **Klimenko D.V., Nikitin A.B., Stroganov A.A., Tsikin I.A.** Design of monolithic microwave integrated circuit X-band switch on GaAs pHEMT. *Radiotekhnika*, 2024, Vol. 88, No. 3, Pp. 92–101. DOI: 10.18127/j00338486-202403-09

11. **Klimenko D.V., Nikitin A.B., Stroganov A.A.** Monolithic integrated circuit of a four-channel switched filter bank for the centimeter band based on GaAs pHEMT technology. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 114–123. DOI: 10.18721/JCSTCS.17311

12. **Osipov A.M.** Limiters on the basis of Shottki diodes. *2007 17th International Crimean Conference – Microwave & Telecommunication Technology*, 2007, Pp. 97–98. DOI: 10.1109/CRMICO.2007.4368642

13. **Smith D.G., Heston D.D., Allen D.L.** Designing high-power limiter circuits with GaAs PIN diodes. *1999 IEEE MTT-S International Microwave Symposium Digest* (*Cat. No.99CH36282*), 1999, Vol. 1, Pp. 329–332. DOI: 10.1109/MWSYM.1999.779486

14. **Pham B.L., Ngueyn D.P., Pham A.-V., Le P.D.** High Power Monolithic pHemt GaAs Limiter for T/R Module. *IEEE Compound Semiconductor Integrated Circuit Symposium* (*CSICS*), 2016, pp. 1–4. DOI: 10.1109/CSICS.2016.7751041

15. **Devlin L.M., Dearn A.W., Beasly P.D.L.** A Monolithic, Dual Channel, 0.5 to 20 GHz Limiter. *2000 30th European Microwave Conference*, 2000, Pp. 1–4. DOI: 10.1109/EUMA.2000.338687

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Denis V. Klimenko
Клименко Денис Валерьевич
E-mail: devklimenko@stc-spb.ru

Aleksandr B. Nikitin
Никитин Александр Борисович
E-mail: alexa_nik@spbstu.ru

Alexander A. Stroganov
Строганов Александр Александрович
E-mail: stroganov187@yandex.ru

Submitted: 17.02.2026; Approved: 23.04.2026; Accepted: 04.05.2026.
Поступила: 17.02.2026; Одобрена: 23.04.2026; Принята: 04.05.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 61–69.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 61–69.

Short message
DOI: https://doi.org/10.18721/JCSTCS.19206
UDC 621.3.049.774.2

A SPIKING NEURAL NETWORK IN 180 NM CMOS

M.M. Pilipko ✉ , *D.V. Morozov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ m_m_pilipko@rambler.ru

Abstract. A hardware implementation of a spiking neural network for classifying font images of letters and numbers has been developed in the 180 nm CMOS technology of Mikron JSC. The layout dimensions are 500 × 550 μm. The supply voltage is 1.0 V. The average power consumption is approximately 0.06 mW. The clock frequency is up to 10 MHz. The neural network contains four layers of 15, 8, 8, and 10 nodes, respectively. The input data and synapse weights have a bit width of 3 bits. A comparison with the results from other studies is performed.

Keywords: CMOS, analog neural network, leaky integrating neuron

Acknowledgements: The integrated circuit was manufactured using funds from the Ministry of Science and Higher Education of the Russian Federation within the framework of the federal project “Training of personnel and scientific foundation for the electronics industry” under the state assignment “Development of a prototyping methodology for electronic component base using domestic microelectronic manufacturing facilities based on MPW service”.

Citation: Pilipko M.M., Morozov D.V. A spiking neural network in 180 nm CMOS. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 61–69. DOI: 10.18721/JCSTCS.19206

© Pilipko M.M., Morozov D.V., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Краткое сообщение
DOI: https://doi.org/10.18721/JCSTCS.19206
УДК 621.3.049.774.2

ИМПУЛЬСНАЯ НЕЙРОННАЯ СЕТЬ ПО ТЕХНОЛОГИИ КМОП 180 НМ

М.М. Пилипко ✉ , *Д.В. Морозов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ m_m_pilipko@rambler.ru

Аннотация. Разработана аппаратная реализация импульсной нейронной сети для классификации изображений шрифта букв и цифр в КМОП-технологии АО «Микрон» 180 нм. Размеры топологии 500 × 550 мкм. Напряжение питания 1,0 В. Средняя потребляемая мощность около 0,06 мВт. Тактовая частота до 10 МГц. Сеть содержит четыре слоя по 15, 8, 8 и 10 узлов соответственно. Входные данные и веса синапсов имеют разрядность 3 бита. Проведено сравнение с результатами других работ.

Ключевые слова: КМОП, аналоговая нейронная сеть, интегрирующий нейрон с утечкой

Финансирование: Производство интегральной микросхемы было выполнено за счет средств Министерства науки и высшего образования Российской Федерации в рамках федерального проекта «Подготовка кадров и научного фундамента для электронной промышленности» по государственному заданию «Разработка методики прототипирования электронной компонентной базы на отечественных микроэлектронных производствах на основе сервиса MPW».

Для цитирования: Pilipko M.M., Morozov D.V. A spiking neural network in 180 nm CMOS // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 61–69. DOI: 10.18721/JCSTCS.19206

Introduction

The development of neural networks that can classify data with high speed and accuracy while consuming small power is an emerging research area. Of particular interest are spiking neural networks (SNNs), in which interactions between artificial neurons are mediated by pulses of a specific amplitude passing through functional analogs of biological junctions – synapses [1–8]. The goal of this work is to develop a four-layer analog SNN topology for classifying font images of letters and numbers, in a 5 × 3 pixel format, characterized by a small number of neural nodes and a simple neuron structure. In this paper, a hardware realization of a SNN is developed in the 180 nm CMOS technology of Mikron JSC. Simulation results taking into account the parasitic parameters of the circuit layout and pads, i.e., post-layout simulation results, are presented.

Spiking neural network

An SNN can be used in image classification. An analog neural network is potentially more efficient than a digital one. In this work, the neural network has 15 nodes in the input layer, 10 neurons in the output layer, and also has two hidden layers with 8 neurons each (Fig. 1, *a*).

Fig. 1*b* shows the main blocks of the SNN chip. The chip consists of: input data register, control circuit, spike generators, synapse register, first and second hidden layers, output layer, and counters with output register. The input data, as well as the synapse weights, are fed to a register in parallel form. The input layer generates spikes, whose number is proportional to the input value. Synapses transmit these spikes to neurons in the next layer. The neuron's capacitor is charged by the spikes,

© Пилипко М.М., Морозов Д.В., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

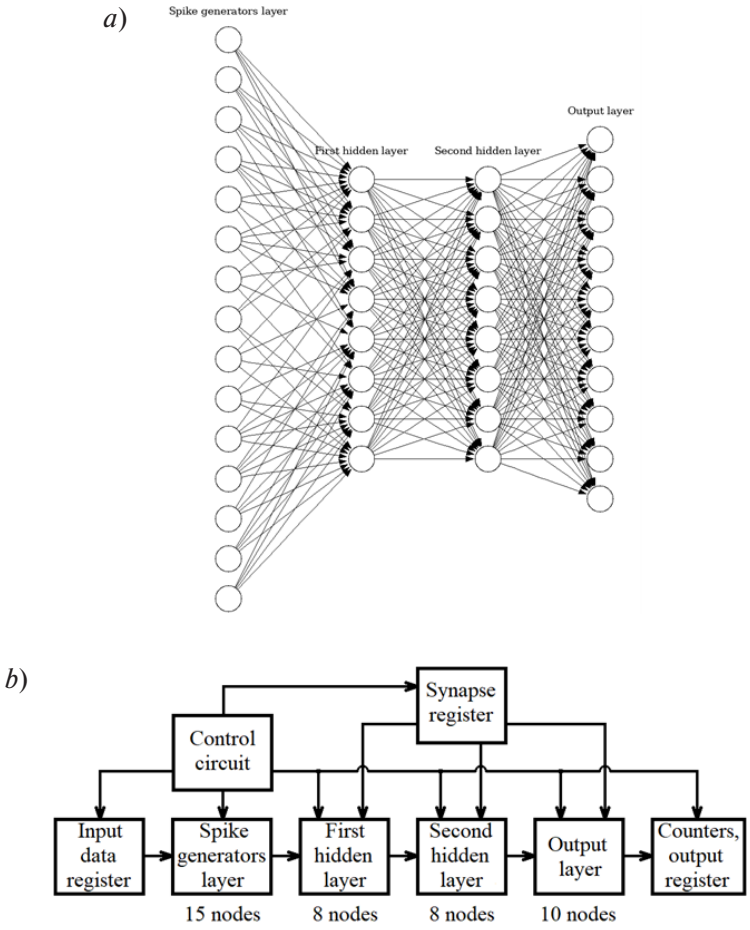


Fig. 1. SNN

with the current proportional to the synapse weight. When a threshold is exceeded, the neuron generates ("fires") its own spike, which is passed on to the next layer, while the capacitor is reset. Neurons are implemented using the leaky integrate-and-fire (LIF) neuron model. Spikes from neurons in the output layer are counted by counters. The counter codes are the network's output signal. The decision corresponds to the node that has produced the maximum code. The input data and the synapse weights have a 3-bit resolution. Silence mode, synapse weights setting mode, data input/output mode, and operating mode are supported. The working cycle including input data writing, network operation, and data output, takes 36 clock periods. Default synapse weights and default input data can be set for control. The comparator thresholds in each internal layer and in the output layer can be controlled. The neuron charge leakage rate can be adjusted. The output of one of the neurons is routed from each layer to control spike generation.

Examples of the synapse circuit are presented in [2, 6–8]. The synapse circuit designed in this work is shown in Fig. 2, *a*. In accordance with Fig. 1, the input of each neuron is connected to eight synapses. The spike "SPK" is transmitted via the synapse. The weight of the synapse is set using the values "W<0:2>" that are stored in a row of synapse register (Fig. 1, *b*). The output node "MEM" is used to connect neuron capacitor. The data "D<0:2>" on MOSFETs' gates of single-finger M1, two-finger M2, and four-finger M3 determines charge current of the neuron capacitor.

The neuron circuit designed in this work is shown in Fig. 2, *b*. The circuit consists of three parts: comparator on MOSFETs M1–M11; analog memory element on MOSFETs M12–M16, neuron capacitor C1, and inverter I9; and logic part I1–I8. The signal "LEAK" sets the rate of charge leakage

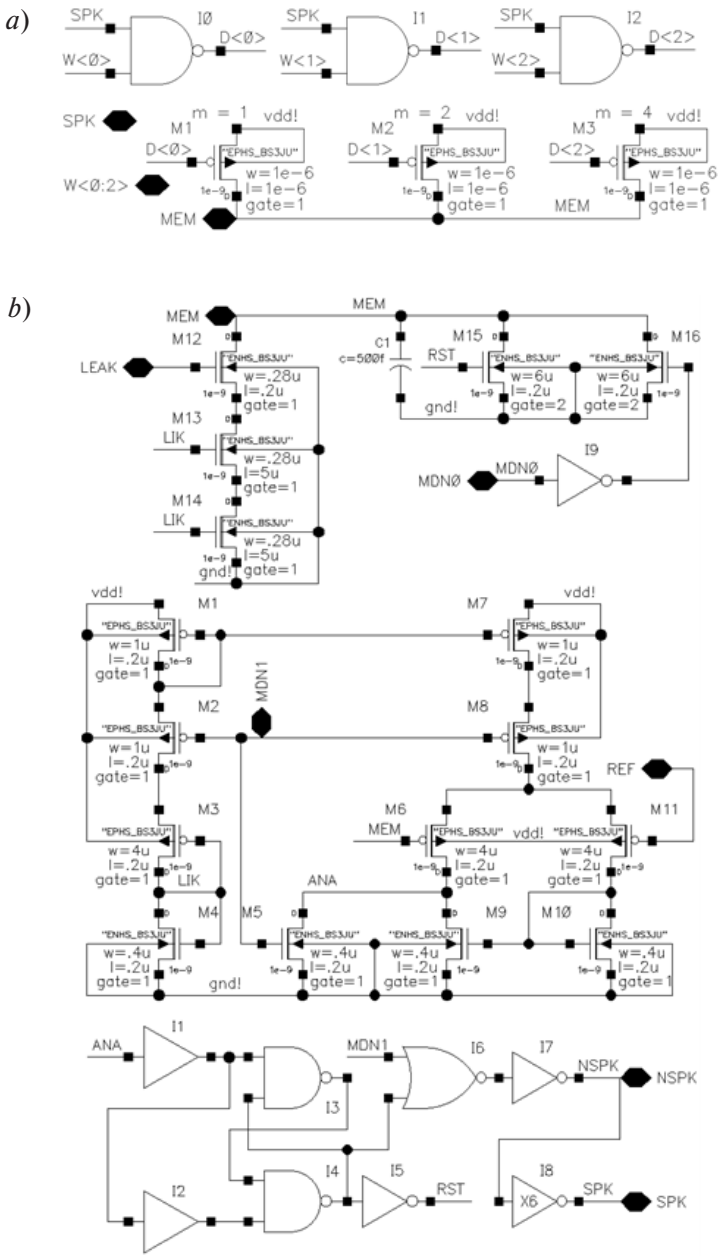


Fig. 2. Synapse and neuron circuits

of the neuron capacitor. The signal “MDN1” enables the comparator operation during operating mode of the SNN and disables the comparator in other modes. The comparator compares the signals “REF” and “MEM”, generates its output signal “ANA”, which is the input signal to the logic. Buffer I1 consists of four series-connected identical CMOS inverters. Buffer I2 consists of six series-connected CMOS inverters with different transistor lengths and defines the pulse width of the output spike. The NANDs I3 and I4 form an SR latch, while CMOS inverter I5 produces reset signal “RST” for discharging the neuron capacitor C1. The signal “MDN0” enables the discharge of the neuron capacitor C1 in silence mode. The signal “SPK” is the neuron’s output, while the signal “NSPK” provides the ability to control data in layers of the SNN using measuring equipment. The CMOS inverter I8 uses six-finger MOSFETs to increase neuron’s load capacity. The signal “LIK”, formed in the comparator, is an additional biasing signal for analog memory element.

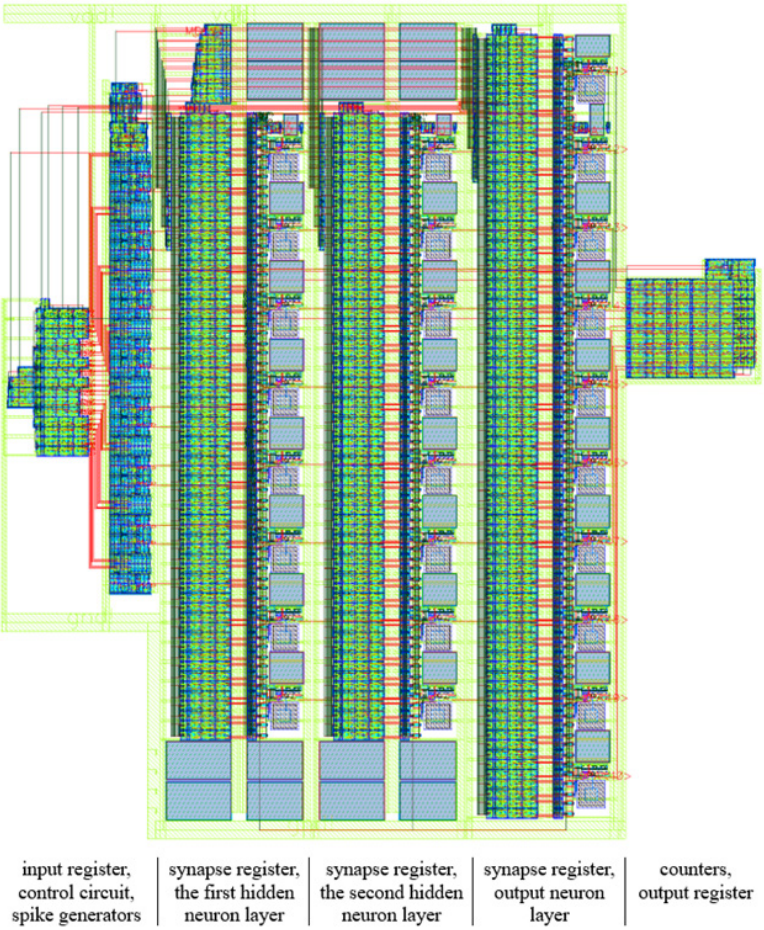


Fig. 3. SNN circuit layout

The layout of the SNN circuit is shown in Fig. 3 and has dimensions 500 × 550 μm. The layout consists of columns representing some main parts, namely input data register, control circuit and spike generators, synapse registers, neurons of the first and second hidden layers, neurons of the output layer, and counters with the output register in accordance with Fig. 1, *b*.

The supply voltage of Mikron JSC 180 nm CMOS technology is 1.8 V, but at this voltage the SNN circuit has high-power consumption, ca. 0.5 mW. It is possible to decrease power consumption by using supply voltage of 1.0 V. At this voltage the SNN circuit still works correctly while consuming only 0.06 mW.

Fig. 4 shows the operation of a neuron. The signal "CLK" represents the input spikes from previous neural network layer. The reference level "REF" is equal to 250 mV (dot-dashed line). There are two signals of neuron memory shown at different leakage control voltages that are equal to 0.1 V and 1.0 V. For "MEM_0.1" (dashed line), the leakage of the neuron is less than for "MEM_1.0" (continuous line). If the reference level is exceeded by the value in the neuron's memory, a spike "SPK" is formed.

Fig. 5 shows the process of writing the input data into 15 neurons of the SNN input layer as well as generation of spikes as a results of post-layout simulation. Signals "MODE<1>" and "MODE<2>" define SNN working mode, at MODE<1> = 0 and MODE<2> = 1, the input data writing is performed. The input data has a 3-bit resolution and is shown in Fig. 5 in decimal representation as "CODE" from 0 to 7. The input data writing is performed at a fall of the signal pulse "CLK<2>". At MODE<1> = = 0 and MODE<2> = 0, the neural network is in silence mode. The outputs of spike generators in input layer are indicated from "SP_1<1>" to "SP_1<15>". In operating mode at MODE<1> = 1 and

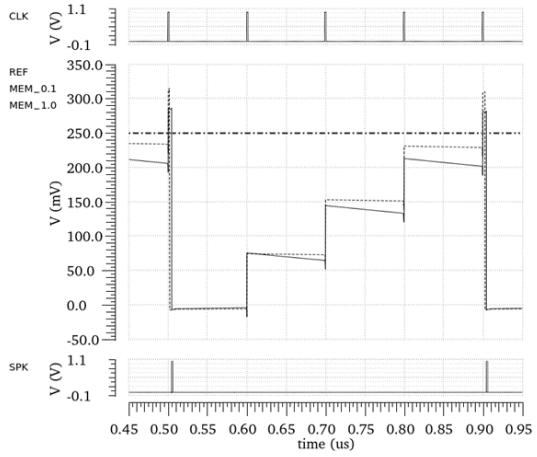


Fig. 4. Neuron operation

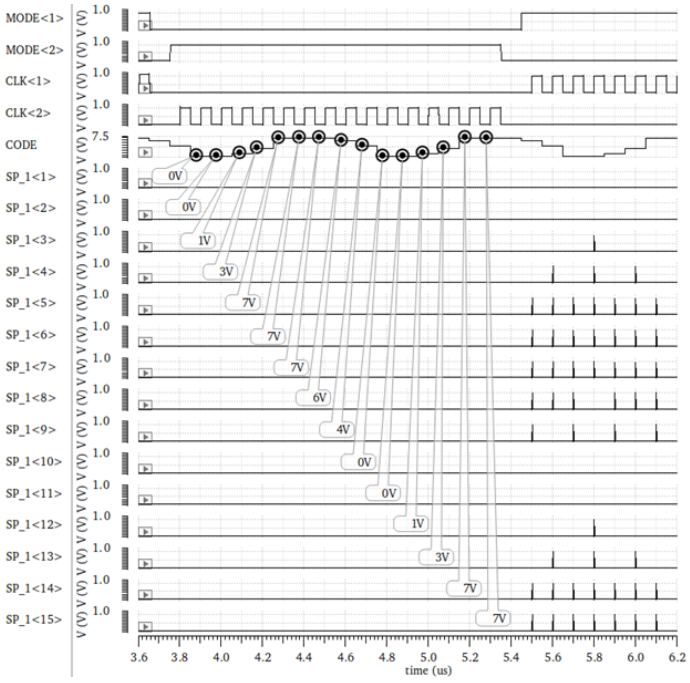


Fig. 5. The input data writing and spikes generation

MODE<2> = 0, the input layer generates spikes (their number can vary from 0 to 7) at the rise of the signal pulse “CLK<1>”. The number of spikes is equal to the input value of “CODE”. Synapses transmit these spikes to neurons in the next layer.

Fig. 6, *a* shows spikes generation in four layers of the neural network as the results of post-layout simulation. As seen, in operating mode at MODE<1> = 1 and MODE<2> = 0, the input layer generates six spikes on “SP_1<1>”, the first hidden layer – three spikes on “SP_2<1>”, the second hidden layer – two spikes on “SP_3<1>”, and the output layer – two spikes on “SP_4<1>”. To check the data in SNN layers using standard measuring equipment, signals “SP_1_pulse”, “SP_2_pulse”, “SP_3_pulse”, and “SP_4_pulse” are formed and have detectable pulse durations ca. 0.05 μs similar to that of “CLK<1>”.

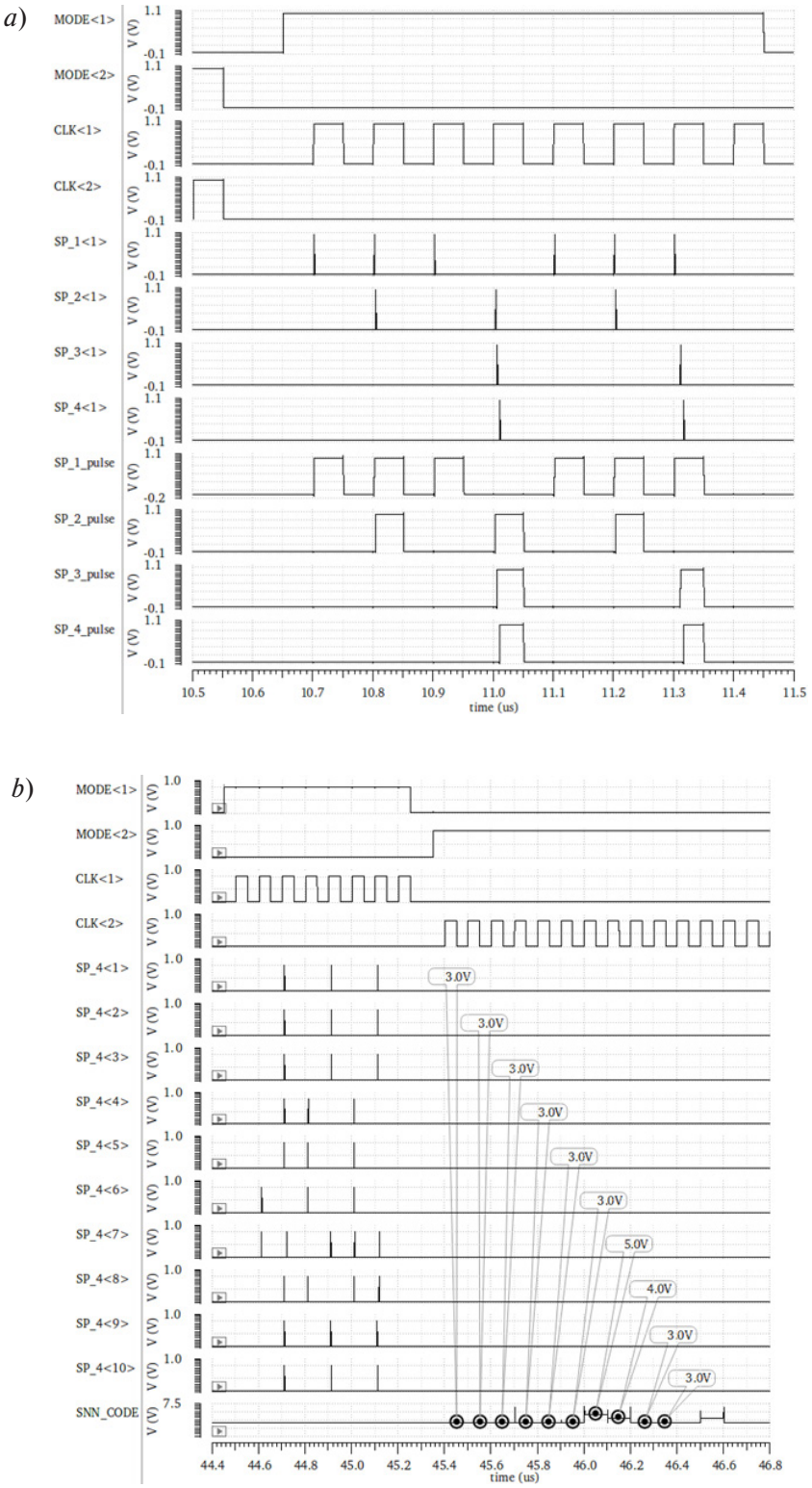


Fig. 6. Spikes generation in four layers and SNN result output

Fig. 6*b* shows the process of reading the output data out of 10 neurons of the SNN output layer as the results of post-layout simulation. The spikes of the output layer are indicated from "SP_4<1>" to "SP_4<10>". In operating mode at MODE<1> = 1 and MODE<2> = 0, the output layer generates spikes from 0 to 7 at the rise of the signal pulse "CLK<1>". At MODE<1> = 0 and MODE<2> = 0,

the neural network is in silence mode. The output data reading is performed at MODE<1> = 0 and MODE<2> = 1. The output data has a 3-bit resolution and is shown in Fig. 6*b* in decimal representation as "SNN_CODE", i.e., the values can be from 0 to 7. The output data reading is performed at the fall of the signal pulse "CLK<2>". As seen, the value of "SNN_CODE" is proportional to the number of spikes.

Main characteristics of the realized SNN are summarized in Table 1 in comparison with the results published in [2]. Complexity/Area and Complexity/Power are near to those in [2].

Table 1

Comparison of results

Main characteristics	This work	[2]
CMOS technology	180 nm	180 nm
Architecture	Analog	Analog
Classifier type	SNN	SNN
Neuron Model	LIF	LIF
Area	0.275 mm^2	3.6 mm^2
Power consumption	0.06 mW	1.06 mW
Complexity [Total number of weights]	208	3311
Complexity/Area	756	920
Complexity/Power	3467	3123

Conclusion

The hardware realization of the SNN in the 180 nm CMOS technology of Mikron JSC has been discussed. In comparison with the results published in [2], this work provides competitive values of Complexity/Area and Complexity/Power characteristics.

REFERENCES

1. **Asghar M.S., Arslan S., Kim H.W.** Low Power Spiking Neural Network Circuit with Compact Synapse and Neuron Cells. *2020 International SoC Design Conference* (*ISOCC*), 2020, Pp. 157–158. DOI: 10.1109/ISOCC50952.2020.9333105

2. **Asghar M.S., Arslan S., Kim H.W.** A Low-Power Spiking Neural Network Chip Based on a Compact LIF Neuron and Binary Exponential Charge Injector Synapse Circuits. *Sensors*, 2021, Vol. 21, No. 13, Art. no. 4462. DOI: 10.3390/s21134462

3. **Skrebenkov D.I., Budanov D.O.** Hardware Implementation of Neural Networks. *2024 International Conference on Electrical Engineering and Photonics* (*EExPolytech*), 2024, Pp. 99–102. DOI: 10.1109/EExPolytech62224.2024.10755531

4. **Skrebenkov D.I., Budanov D.O.** Methods for Reducing Power Consumption in Hardware Implemented Neural Networks. *2025 International Conference on Electrical Engineering and Photonics* (*EExPolytech*), 2025, Pp. 1–4. DOI: 10.1109/EExPolytech66949.2025.11252131

5. **Mavrin I.A., Ryndin E.A., Andreeva N.V.** Razrabotka analogovoi impul'snoi neironnoi seti [Development of an analog pulse neural network]. *XI Nauchno-prakticheskaia konferentsiia s mezhdunarodnym uchastiem "Nauka nastoiashchego i budushchego" dlia studentov, aspirantov i molodykh uchenykh* [*XI Scientific and Practical Conference with International Participation "Science of the Present and Future" for students, postgraduates and young scientists*], 2023, Vol. 1, Pp. 114–118.

6. **Udovichenko V., Krokhmanskii V., Ryndin E.A.** *Topologiia iskusstvennogo sinapsa dlia avtomatizirovannogo proektirovaniia neironnykh setei* [*Artificial synapse topology for automated design of neural networks*], Available: https://events.priem.etu.ru/api/open/nnb/stand-file?id=85844 (Accessed 24.06.2026)

7. **Ryndin E.A., Andreeva N.V.** *Integral electronic CMOS synapse*. Russian Federation Patent, No. RU 2808951 C1 (2023).

8. **Ryndin E.A., Andreeva N.V.** *Integrated electronic synaptic CMOS element*. Russian Federation Patent, No. RU 2836650 C1 (2024).

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Mikhail M. Pilipko
Пилипко Михаил Михайлович
E-mail: m_m_pilipko@rambler.ru
ORCID: https://orcid.org/0000-0003-3813-6846

Dmitry V. Morozov
Морозов Дмитрий Валерьевич
E-mail: dvmorozov@inbox.ru
ORCID: https://orcid.org/0000-0003-3403-0120

Submitted: 21.04.2026; Approved: 22.05.2026; Accepted: 28.05.2026.
Поступила: 21.04.2026; Одобрена: 22.05.2026; Принята: 28.05.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 70–80.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 70–80.

Simulations of Computer, Telecommunications and Control Systems

Моделирование вычислительных, телекоммуникационных и управляющих систем

Research article
DOI: https://doi.org/10.18721/JCSTCS.19207
UDC 004.925

ANALYSIS OF THE INFLUENCE OF WEIGHTING COEFFICIENTS IN A SIGNIFICANCE METRIC FOR 3D SCENE OBJECTS IN PRELIMINARY RENDERING OPTIMIZATION SETUP

V.G. Chernyi ✉, *M.V. Bolsunovskaya*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ vitaly.g.cherny@gmail.com

Abstract. This paper presents an experimental analysis of the behavior of a significance metric for 3D scene objects used for preliminary rendering optimization setup in the Unity environment. The influence of three object characteristics is considered, from which the final significance metric is formed: geometric complexity (polygon count), spatial size, and the object's potential as an occluder, calculated based on visibility analysis performed using ray casting. Combining these features with adjustable weights makes it possible to obtain a unified estimate of the potential importance of an object and its impact on rendering during the configuration of optimization approaches. The results show that geometric complexity provides strong separation of objects into groups, assigning 22.6% of the objects in the experimental scene to the high-significance category. The balanced configuration and the dominance of other coefficients individually demonstrate lower differentiation, yet accurately identify the most complex objects intentionally placed in the scene, while classifying over 70% of the remaining models as low-significance. Regardless of the chosen configuration, the method correctly identifies candidates clearly suitable for further optimization.

Keywords: computer graphics, 3D scene analysis, visibility analysis, rendering optimization, Unity engine

Citation: Chernyi V.G., Bolsunovskaya M.V. Analysis of the influence of weighting coefficients in a significance metric for 3D scene objects in preliminary rendering optimization setup. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 70–80. DOI: 10.18721/JCSTCS.19207

© Chernyi V.G., Bolsunovskaya M.V., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19207
УДК 004.925

АНАЛИЗ ВЛИЯНИЯ КОЭФФИЦИЕНТОВ МЕТРИКИ ЗНАЧИМОСТИ ОБЪЕКТОВ 3D-СЦЕНЫ ДЛЯ ПРЕДВАРИТЕЛЬНОЙ НАСТРОЙКИ ОПТИМИЗАЦИИ РЕНДЕРИНГА

В.Г. Чёрный ✉ , *М.В. Болсуновская*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ vitaly.g.cherny@gmail.com

Аннотация. В работе проведен анализ поведения метрики значимости объектов трехмерной сцены и влияния на нее коэффициентов в рамках метода для предварительной настройки оптимизации рендеринга в среде Unity. Рассматривается влияние трех характеристик объекта, из которых формируется итоговая метрика значимости: геометрическая сложность (количество полигонов), пространственный размер и потенциал объекта в качестве окклюдера, вычисляемый на основе анализа видимости, выполняемого с помощью запуска лучей. Объединение этих признаков с настраиваемыми весами позволяет получить единую оценку потенциальной важности объекта и его влияния на рендеринг при настройке оптимизационных подходов. Результаты показывают, что геометрическая сложность обеспечивает сильное разделение объектов по группам, относя 22,6% объектов на экспериментальной сцене к группе высокой значимости. Сбалансированный запуск и доминирование других коэффициентов по отдельности демонстрируют меньшую дифференциацию, но точное определение заранее встроенных в сцену самых комплексных объектов, при этом относя более 70% остальных моделей к малозначащим. Независимо от выбранной конфигурации, метод корректно выявляет кандидатов, явно подходящих для дальнейшей оптимизации.

Ключевые слова: компьютерная графика, анализ 3D-сцен, анализ видимости, оптимизация рендеринга, движок Unity

Для цитирования: Chernyi V.G., Bolsunovskaya M.V. Analysis of the influence of weighting coefficients in a significance metric for 3D scene objects in preliminary rendering optimization setup // Computing, Telecommunications and Control. 2026. T. 19, № 2. C. 70–80. DOI: 10.18721/JCSTCS.19207

Introduction

Nowadays, modern 3D interactive applications are developed using 3D engines such as Unity and Unreal Engine. Despite their popularity in the gaming industry, these engines are also widely used in projects related to education, industrial simulations, and the creation of digital twins [1–5]. As technology evolves, the complexity of such projects continues to grow. 3D scenes require more computations and contain increasingly detailed objects, while still demanding smooth performance and fast user feedback [6–7].

To meet these requirements, developers employ various optimization techniques. A large number of optimization tools are already integrated into modern 3D engines, but they require proper configuration to be effectively applied. The need to maintain a balance between visual quality and performance makes research on the application of optimization techniques in 3D engines both relevant and essential [8].

Among the most commonly used in-engine optimization approaches are Level of Detail (LOD) and occlusion culling systems. LOD techniques adjust the geometric complexity of objects based on

© Чёрный В.Г., Болсуновская М.В., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

their distance from the camera. Each object contains multiple levels of detail that are switched according to predefined thresholds. Occlusion culling, on the other hand, reduces the rendering workload by excluding objects that are not visible to the camera due to occlusion by other objects (occluders), or simply because they are outside the field of view [9–12].

All things considered, both of these approaches, in their integrated implementations, require configuration. Such configurations heavily rely on manual decisions made by developers. To achieve efficient use of these tools, developers must determine which objects are important, which can be simplified, and which can serve as occluders for surrounding objects. In large and complex 3D scenes containing objects of varying quality, these decisions are often based on heuristic assumptions. One way to simplify this process and increase its efficiency is to introduce a formalized evaluation framework that helps analyze which objects require optimization techniques.

Existing research primarily focuses on improving real-time rendering optimization algorithms [13–15] or visibility techniques [16–17], while less attention is given to the analysis of objects targeted for optimization during the scene configuration stage. In particular, the combined effect of geometric, spatial, and visibility-related features on object importance remains insufficiently studied.

Related studies also show that the problem of identifying visually or functionally important elements of a 3D scene has already been addressed from different perspectives, including mesh saliency [18–22], viewpoint entropy [23], learning-based view selection [24], and perceptual rendering [25]. These works confirm that scene content can be differentiated by importance, visibility relevance, or perceptual contribution. However, such approaches are mainly focused on perception, view planning, or runtime rendering behavior rather than on the preliminary configuration of built-in optimization tools in 3D engines. In this context, the formalized evaluation of scene objects for editor-based optimization support remains insufficiently explored.

To address this gap, this work considers an object significance metric as a key factor in scene analysis. This unified metric is used to evaluate objects based on their features, such as geometric complexity, spatial size, and occlusion potential calculated through ray-based analysis.

This research investigates the behavior of these calculated metrics under different parameter configurations. The main objective is to analyze how different values of feature weights affect the analysis results and optimization-related decisions, such as object classification and occluder selection.

The key contributions of this paper are as follows:

- analysis of the influence of object properties on significance-based evaluation;
- experimental study of object classification and occluder selection under different coefficient settings.

The proposed approach is implemented and evaluated within the Unity engine using experimental 3D scenes containing different amounts of primitive and low-poly objects. The experimental results provide insights into the behavior of significance-based scene analysis and demonstrate the role of weighting coefficients in this approach.

Proposed approach

The developed method is based on the preliminary analysis of 3D scene objects through the formation of a unified significance metric. The objective is to provide a formalized representation of each object's importance, which can be used for subsequent automated or manual optimization decisions, such as level of detail assignment and occluder selection.

The proposed approach operates on a set of object data collected from a 3D scene and evaluates each object independently by analyzing its predefined properties. These properties reflect how an object contributes to rendering performance and the visual perception of space in a 3D scene. While the contribution of many object properties deserves further investigation, this study primarily considers three features.

1. Geometric complexity

Geometric complexity is defined as the number of polygons (triangles) that constitute the analyzed object. This value reflects the computational cost of rendering the object and is commonly used in LOD systems, since one of the most straightforward ways to reduce rendering cost is to decrease the number of processed polygons.

2. Spatial size

Spatial size represents the physical extent of an object within the scene. It is computed based on the dimensions of the object's bounding volume. By considering spatial size, it becomes possible to evaluate how much space an object occupies. Larger objects are more likely to influence visibility and, therefore, may act as occluders that block other objects. However, spatial size alone does not capture the full scene context. For instance, a large object located at the boundary of the scene may not function as an occluder, but rather as a candidate for LOD configuration.

3. Occlusion potential

Occlusion potential is calculated using a ray-based analysis. A set of rays is generated from a predefined number of source points around the analyzed object. Intersections with other objects, excluding the target itself, are recorded. If the number of intersections exceeds a specified threshold, the object can be marked as an occluder (for example, by assigning the "Occluder Static" flag in the Unity engine). Occlusion potential reflects the likelihood of an object blocking visibility by analyzing its surrounding environment, providing an approximation of its contribution to occlusion culling.

To combine the considered object characteristics into a unified value, a significance metric is introduced:

$$V(o) = k1 \cdot C(o) + k2 \cdot S(o) + k3 \cdot H(o),$$

where $V(o)$ is the significance value of object o; $C(o)$ is the normalized geometric complexity of object (number of triangles); $S(o)$ is the normalized spatial size (object's bounding box volume); $H(o)$ is the normalized occlusion potential estimated by using ray-based analysis; $k1$, $k2$, and $k3$ are weighting coefficients that define the contribution of each feature and satisfy the condition $k1 + k2 + k3 = 1$.

Prior to aggregation, all object feature values are normalized to the range [0, 1] using min-max normalization. This ensures comparability and interpretability between values of different scales.

An additive aggregation was chosen instead of multiplicative model in order to maintain balance between features, avoiding possible suppression effect caused by low-valued components. Using an additive model also simplifies sensitivity analysis, since it was important to see how different coefficient setups influence resulting metric. The application of multiplicative models is considered as a potential direction for future work.

The significance metric produces a scalar value representing the overall importance of an object within the scene. The weighting coefficients can be adjusted to emphasize different aspects of object importance, such as rendering cost (geometric complexity), spatial size, or occlusion capability.

The choice of a linear additive model for significance metric is also motivated by several practical considerations derived from 3D application development. First, linear aggregation ensures computational transparency during scene preprocessing, allowing easy and fast result interpretation. Second, the additive form provides direct traceability: the influence of each coefficient can be immediately understood and tuned. Third, the metric is supposed to be used in context of scene preprocessing, where rapid iteration and predictable behavior are in priority. It is also important that chosen model does not require high computational costs and is easy to integrate in project analysis pipeline. The model may get more complex based on discovered coefficient influence and other further research.

By incorporating the computed significance value into 3D scene analysis, more informed decisions can be made during the configuration of optimization methods. First, the metric can be applied

to occluder selection. A high significance value may indicate both strong occlusion potential and substantial spatial presence, making such objects more likely to block visibility and serve as effective occluders.

To conduct controlled experimental studies of the metric behavior, several configurable parameters are defined prior to analysis. These include the number of rays and their sources, threshold values for occluder detection and classification, as well as the weighting coefficients.

In the context of this study the selection of weighting coefficients is performed empirically to define representative configurations for analysis. Flexible coefficient tuning is supported, allowing adaptation to different 3D scene properties and optimization goals. Several configurations are considered, including balanced weighting and setups based on individual dominance of each feature. The choice of coefficients may depend on the type of the analyzed scene. For example, geometry-dominant analysis can be beneficial for scenes containing big amount of highly detailed objects. These experimental configurations are designed to investigate this idea. The derivation of optimal coefficient values for specific scene types is a prospect for future research.

The proposed framework serves as a basis for analyzing the influence of different weighting coefficient combinations on optimization decisions, which is further explored in the experimental section.

Research methods

The purpose of the research is to investigate the influence of characteristic weighting on the behavior of the proposed significance metric and its impact on optimization configuration decisions in 3D scenes. Following research methods were employed:

1. Controlled experiment. Four fixed configurations of coefficients were defined and tested on the same 3D scene to check the effect of each configuration in the same context.

2. Comparative analysis. The analysis finishes with object classification across three significance groups (low, medium and high). The results of these classifications were compared for each configuration.

3. Mean significance metric value was calculated for each configuration to check overall tendency of metric.

4. Five large-scale objects were intentionally placed in the scene to verify that any configuration will constantly identify evidently important objects.

Experimental setup

The experiments were conducted using a pre-built 3D scene created in the Unity engine editor. The test scene consists of a large number of static simple building models (up to 1000), representing an urban-like environment with varying object scales and high density. Most of the objects are not unique, as they are instances of seven base building models. These base objects differ in spatial size and geometric complexity. Such a configuration ensures diverse object combinations within the scene, making the task of identifying significant objects more challenging. Among all objects, five of them have increased size to imitate evident occluders. In addition, they differentiate in color. Fig. 1 shows model examples, while Fig. 2 demonstrates scene setup from the top-down view in Editor Mode.

Since the main goal of the study is to investigate the role of weighting coefficients in the computation of the significance metric, the parameters related to occlusion potential analysis (number of rays, number of ray sources, and occluder threshold) were predefined and kept constant throughout all experiments. Four experimental cases were defined, each corresponding to a different configuration of weighting coefficients.

The specific values of the weighting coefficients used in the experiments are defined as follows:

1. Balanced configuration: $k1 = 0.33$, $k2 = 0.33$, $k3 = 0.33$;
2. Geometry-dominant configuration: $k1 = 0.6$, $k2 = 0.2$, $k3 = 0.2$;

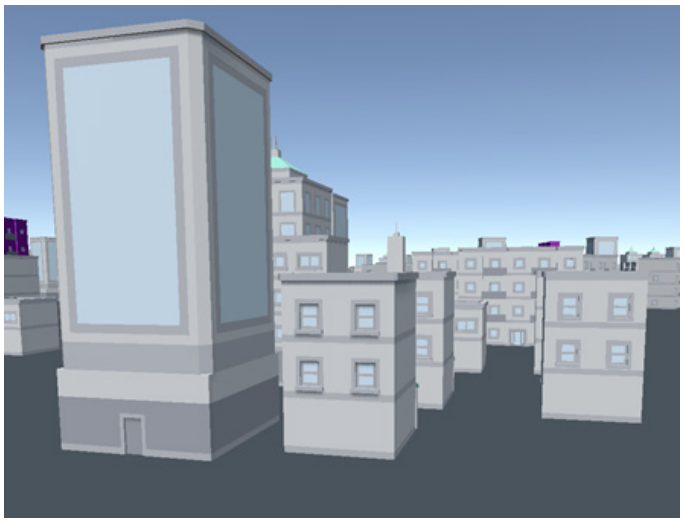


Fig. 1. Object examples

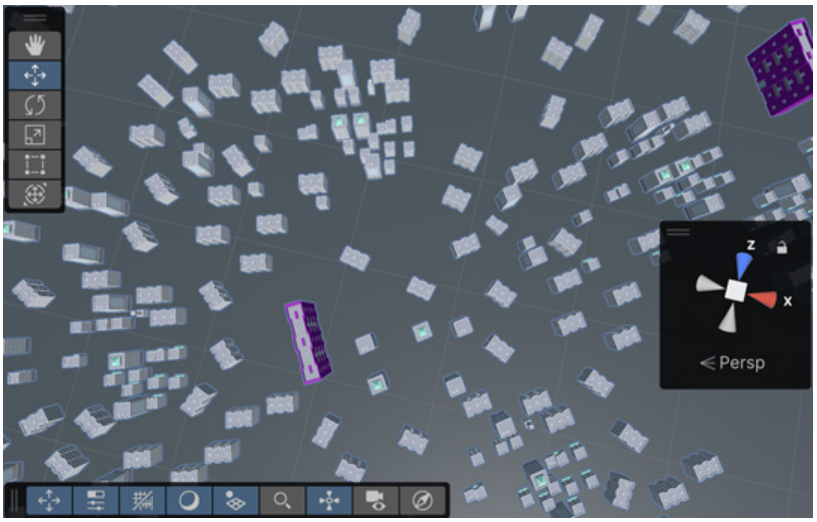


Fig. 2. Scene top-down view

3. Occlusion-dominant configuration: $k1 = 0.2$, $k2 = 0.2$, $k3 = 0.6$;
4. Size-dominant configuration: $k1 = 0.2$, $k2 = 0.6$, $k3 = 0.2$.

These configurations were selected to explicitly analyze the influence of each feature on the resulting significance metric. The first experiment is based on a balanced configuration, where all coefficients have equal importance. The second experiment uses a geometry-dominant configuration. By increasing the importance of geometric complexity, the analysis prioritizes objects with a high polygon count and, consequently, higher rendering cost. The third experiment focuses on occlusion potential, where the main objective is to analyze differences in occluder selection across the three configurations. Fourth experiment uses spatial size of the object.

The evaluation of the experiments is based on the distribution of objects across significance ranges and the number of selected occluders. Three significance levels are introduced based on the value of $V(o)$: low (≤ 0.33), medium ($0.33 < V(o) \leq 0.66$) and high (> 0.66). These thresholds correspond to an equal partitioning of the normalized value range. A simple partitioning based on three groups ensures

consistent comparison between different configurations and keeps analysis results easily interpretable. The results are compared to determine how different weighting schemes influence optimization decisions.

The experiments were performed within the Unity editor environment, inside a project containing the implementation of the proposed approach along with the required 3D assets. The analysis tool is implemented as a custom editor extension, enabling automated processing of the scene. The analysis is performed outside of runtime, as its purpose is to support configuration and optimization rather than real-time execution.

Results and discussion

The results of the experimental study are presented in Tables 1 and 2.

Table 1

Distribution of objects by classification levels

Configuration	Coefficient values	Low significance group	Medium significance group	High significance group
Balanced	$k1 = 0.33$, $k2 = 0.33$, $k3 = 0.33$	701	294	5
Geometry	$k1 = 0.6$, $k2 = 0.2$, $k3 = 0.2$	496	278	226
Occlusion	$k1 = 0.2$, $k2 = 0.2$, $k3 = 0.6$	700	295	5
Spatial size	$k1 = 0.2$, $k2 = 0.6$, $k3 = 0.2$	986	9	5

Table 1 shows the distribution of objects across significance-based classification levels for different weighting configurations of the metric. The classification is derived from normalized values of the significance metric $V(o)$, which combines geometric complexity, spatial size, and occlusion potential.

It is also important to state again that experimental scene contained five intentionally placed large-scale building models which were adjusted manually to serve as primary occluders. Across all experiment configurations these five objects consistently received the highest significance scores. These objects were always correctly classified as high significance objects, while also marked as occluders by occlusion analysis. Shown invariance to coefficient differences demonstrates that the analysis exhibits robustness in identifying most evident and important objects for optimization, regardless of the emphasis placed on different coefficients.

In the balanced configuration, where all features contribute equally, only five objects are assigned to the high-significance group, which are evident occluders. Most of the models are classified as low-significance objects. Considering the context of experimental scene, this result demonstrates that in case of non-existing domination among the coefficients, only the high-profile objects are treated as significant ones.

When geometric complexity is prioritized, the distribution shifts toward a higher number of significant objects. The share of high-significance objects increases, while the number of low-significance ones decreases compared to other configurations. This indicates that geometric complexity provides a clearer distinction between objects, as it introduces a stronger contrast between simple and complex models.

In contrast, the occlusion-dominant configuration results in a noticeably skewed distribution, similar to balanced configuration, where most objects fall into the low-significance group. This suggests that occlusion potential alone does not provide sufficient differentiation in dense scenes, as many objects demonstrate similar visibility behavior under ray-based analysis.

The most pronounced effect is observed in the size-dominant configuration. In this case, almost all objects are classified as low significance, with only a small number identified as highly significant. It should be noted that several large-scale objects were intentionally included in the scene as reference structures. These objects are correctly classified as highly significant, while the remaining objects form a nearly homogeneous group. This behavior indicates that spatial size alone is not sufficient for reliable object importance evaluation in complex environments.

Table 2 presents the average value of the significance metric for each experimental configuration. The mean value is computed by averaging the significance metric values for each previously analyzed object in the scene. The results reflect the tendency of the metric on current scene. Higher values indicate a shift towards higher significance among the objects, while lower values show that most of the scene consists of less significant 3D models.

The geometry-dominant configuration produces the highest mean value, reflecting a stronger differentiation between objects. The balanced configuration demonstrates moderate values, while the occlusion- and size-dominant configurations result in lower averages, indicating weaker separation between objects.

Table 2

Average values of the significance indicator for each experimental configuration

Configuration	Mean $V(o)$
Balanced	0.23
Geometry	0.31
Occlusion	0.23
Spatial size	0.18

Overall, the results demonstrate that the weighting coefficients of the significance metric influence object classification. Among the considered features, geometric complexity provides the most effective basis for differentiation, while spatial size and occlusion potential, when used independently, show limited capability in distinguishing object importance in dense scenes.

While direct comparison with existing methods is hard to perform due to different evaluation metrics and goals, Table 3 provides a qualitative comparison of the proposed approach with existing techniques mentioned in introduction as related work.

Additionally, in the context of comparing approach discussed in this work with other approaches, it is important to note that recent advances in machine learning, including LLM and vision-based AI systems, provide powerful tools for semantic scene understanding, object importance estimation and image analysis. Those approaches can address similar problems by learning perceptual or significance directly from provided data.

However, the proposed method differs in scope and application context. First, it is meant to be used in the preprocessing stage within 3D engine workflow, where low computational overhead and controllability are very important. Calculation of metric does not require additional training data or any integration outside of engine tools. Second, the proposed approach provides easy interpretability, since each feature can be directly controlled through weighting coefficients. It allows developers to manage and modify analysis process based on their optimization goals and current 3D scene, which can be difficult to achieve with black-box models.

Table 3

Qualitative comparison of the proposed approach with existing methods

Criterion	Mesh saliency [18]	Viewpoint selection [24]	Object-space VSD [16]	Proposed approach
Intended use	Perceptual saliency for mesh simplification and compression	View set selection for synthetic training data generation	Global visible surface determination (object space)	Preliminary in-engine optimization configuration (occluder selection)
Output type	Saliency score per vertex or per face	Set of camera poses	Binary classification (visible/invisible per triangle)	Unified significance metric per object
Requires 3D engine	No	No	No	Yes
Requires training data	Yes (RPCA-generated ground truth)	Yes (NYUDv2)	No	No
Operates on	Individual faces (patch-based)	Rendered images	Individual triangles (per ray)	Whole 3D object
Can affect scene/ objects	No	No	No	Yes (can mark potential occluders in Unity engine)

Conclusion

This paper presents an experimental study of a significance-based approach to object evaluation in 3D scenes within the Unity engine. The proposed metric combines geometric complexity, spatial size, and occlusion potential into a unified scalar value that can be used for object classification in the context of rendering optimization, particularly for occluder selection.

The experiments conducted show measurable shifts in object grouping under different configurations. In particular, the geometry-dominant configuration increases the number of high-significance objects from 5 (in balanced case) to 226, while reducing the number of low significance objects from 701 to 496. Spatial size configuration demonstrates high concentration of objects in low significance group (986 out of 1000), indicating reduced differentiation. The different distribution of objects across classification levels is also shown by mean V(o) values, calculated during experimental runs. Geometry-dominant configuration produces highest average value (0.31), while size-dominant configuration produces the smallest (0.18). The results confirm that different weighting schemes affect object classification based on their significance.

Notably, most of the time, high-significance group received only five objects, which were intentionally placed large-scale occluders. The consistency of analysis results in these cases across most of the configurations confirms the robustness of the significance metric in detecting models which definitely influence scene visibility and rendering complexity.

The proposed approach can be integrated into existing project configuration workflows as a preprocessing tool for rapid 3D scene analysis. By providing a formalized and configurable estimate of object importance, the method reduces reliance on manual judgments, which is particularly valuable for large and complex scene set ups with huge amounts of high-detail models with dense placements.

Future work may proceed in several directions. First, more complex benchmark evaluation is required to research influence on metric values in cases which are closer to real application scenarios. Second, results of rendering optimizations configured based on metric evaluation should be investigated. Finally, adaptive coefficient and analysis parameter selection is an important research prospect, since it can possibly reduce the need for manual value tuning.

The obtained results provide basis for further research in scene analysis based on object significance metric. The critical role of parameter selection and coefficient weighting in such methods is also highlighted, marking the problem of optimization technique configuration adaptation for various 3D scenes.

REFERENCES

1. **Koch J., Gomse M., Schüppstuhl T.** Digital game-based examination for sensor placement in context of an Industry 4.0 lecture using the Unity 3D engine – a case study. *Procedia Manufacturing*, 2021, Vol. 55, Pp. 563–570. DOI: 10.1016/j.promfg.2021.10.077

2. **Anikeev A.S.** Generating and modeling 2D terrain from control values using Unity in the C# programming language. *Engineering Journal of Don*, 2020, Vol. 64, No. 4, Art. no. 16.

3. **Krajčovič M., Gabajová G., Matys M., Grznár P., Dulina Ľ., Kohár R.** 3D Interactive Learning Environment as a Tool for Knowledge Transfer and Retention. *Sustainability*, 2021, Vol. 13, No. 14, Art. no. 7916. DOI: 10.3390/su13147916

4. **Helbig C., Becker A.M., Masson T., Mohamdeen A., Sen Ö.O., Schlink U.** A game engine based application for visualising and analysing environmental spatiotemporal mobile sensor data in an urban context. *Frontiers in Environmental Science*, 2022, Vol. 10, Art. no. 952725. DOI: 10.3389/fenvs.2022.952725

5. **Wang Z., Wu G., Boriboonsomsin K., Barth M.J., Han K., Kim B., Tiwari P.** Cooperative Ramp Merging System: Agent-Based Modeling and Simulation Using Game Engine. *SAE International Journal of Connected and Automated Vehicles*, 2019, Vol. 2, no. 2, pp. 115–126. DOI: 10.4271/12-02-02-0008

6. **Geris A., Cukurbasi B., Kilinc M., Teke O.** Balancing performance and comfort in virtual reality: A study of FPS, latency, and batch values. *Software: Practice and Experience*, 2024, Vol. 54, No. 12, Pp. 2336–2348. DOI: 10.1002/spe.3356

7. **Gonakhchian V.I.** Performance model of graphics pipeline for single-pass dynamic 3D scene rendering scheme. *Proceedings of the Institute for System Programming of the RAS*, 2020, Vol. 32, No. 4, Pp. 53–72. DOI: 10.15514/ISPRAS-2020-32(4)-4

8. **Han S., Sander P.V.** Triangle reordering for efficient rendering in complex scenes. *Journal of Computer Graphics Techniques*, 2017, Vol. 6, No. 3, Pp. 38–52.

9. **Hasselgren J., Andersson M., Akenine-Möller T.** Masked software occlusion culling. *Proceedings of High Performance Graphics* (*HPG'16*), 2016, pp. 23–31. DOI: 10.2312/hpg.20161189

10. **Jiang T., Huang S., Yan Z.** Rendering Optimization of Building Information Models Based on Occlusion Culling. *Proceedings of the 2024 4th International Conference on Public Management and Big Data Analysis* (*PMBDA 2024*), 2025, Pp. 358–368. DOI: 10.2991/978-94-6463-656-7_35

11. **Pantazopoulos I., Tzafestas S.** Occlusion culling algorithms: A comprehensive survey. *Journal of Intelligent and Robotic Systems*, 2002, Vol. 35, Pp. 123–156. DOI: 10.1023/A:1021175220384

12. **Ye L., Liu G., Chen G., Li K., Chen Q., Fan W., Zhang J.** 3D Model Occlusion Culling Optimization Method Based on WebGPU Computing Pipeline. *Computer Systems Science and Engineering*, 2023, Vol. 47, No. 2, Pp. 2529–2545. DOI: 10.32604/csse.2023.041488

13. **Lee E.-S., Shin B.-S.** Vertex Chunk-Based Object Culling Method for Real-Time Rendering in Metaverse. *Electronics*, 2023, Vol. 12, No. 12, Art. no. 2601. DOI: 10.3390/electronics12122601

14. **Ding Y., Song Y.** Vision-Degree-Driven Loading Strategy for Real-Time Large-Scale Scene Rendering. *Computers*, 2025, Vol. 14, No. 7, Art. no. 260. DOI: 10.3390/computers14070260

15. **Kiusya Z., Gikunda P., Musumba G.** Adaptive mesh compression algorithm for near real-time rendering of large-scale static scenes. *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications* (*ACDSA*), 2025, Pp. 1–8. DOI: 10.1109/ACDSA65407.2025.11166552

16. **Rossi A., Barbiero M., Scremin P., Carli R.** Robust Visibility Surface Determination in Object Space via Plücker Coordinates. *Journal of Imaging*, 2021, Vol. 7, No. 6, Art. no. 96. DOI: 10.3390/jimaging7060096

17. **Sun Y., Huang Q., Hsiao D.-Y., Guan L., Hua G.** Learning view selection for 3D scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2021, Pp. 14459–14468. DOI: 10.1109/CVPR46437.2021.01423

18. **Nousias S., Arvanitis G., Lalos A., Moustakas K.** Deep saliency mapping for 3D meshes and applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023, Vol. 19, No. 2, Art. no. 71. DOI: 10.1145/3550073

19. **Nousias S., Arvanitis G., Lalos A., Moustakas K.** Mesh saliency detection using convolutional neural networks. *2020 IEEE International Conference on Multimedia and Expo* (*ICME*), 2020, Pp. 1–6. DOI: 10.1109/ICME46284.2020.9102796

20. **Sarem M., Zheng Y., Wang K., Albshlawy L.** An Improved Deep Learning Based RGB-D Saliency Detection Model. *2025 IEEE 18th International Symposium on Embedded Multicore/Many-core Systems-on-Chip* (*MCSoC*), 2025, Pp. 105–112. DOI: 10.1109/MCSoC67473.2025.00027

21. **Ullah I., Jian M., Hussain S., Guo J., Yu H., Wang X., Yin Y.** A brief survey of visual saliency detection. Multimedia Tools and Applications, 2020, Vol. 79, pp. 34605–34645. DOI: 10.1007/s11042-020-08849-y

22. **Lee C.H., Varshney A., Jacobs D.W.** Mesh saliency. *SIGGRAPH'05: ACM SIGGRAPH 2005 Papers*, 2005, Pp. 659–666. DOI: 10.1145/1186822.1073244

23. **Vázquez P.P., Feixas M., Sbert M., Llobet A.** Viewpoint entropy: a new tool for obtaining good views of molecules. *Joint EUROGRAPHICS – IEEE TCVG Symposium on Visualization*, 2002, Pp. 1–6.

24. **Genova K., Savva M., Chang A.X., Funkhouser T.** Learning where to look: Data-driven viewpoint set selection for 3D scenes. *arXiv:1704.02393*, 2017. DOI: 10.48550/arXiv.1704.02393

25. **Patney A., Salvi M., Kim J., Kaplanyan A., Wyman C., Benty N., Luebke D., Lefohn A.** Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics* (*TOG*), 2016, Vol. 35, No. 6, Art. no. 179. DOI: 10.1145/2980179.2980246

26. **Chernyi V.G., Bolsunovskaya M.V.** Automation of object selection for rendering optimization in Unity. *Razvitie intellektual'noi ekonomiki i promyshlennosti na osnove iskusstvennogo intellekta* [*Development of intellectual economy and industry based on artificial intelligence*], 2025, Pp. 755–771. DOI: 10.18720/IEP/2025.3/36

27. **Meister D., Ogaki S., Benthin C., Doyle M.J., Guthe M., Bittner J.** A Survey on Bounding Volume Hierarchies for Ray Tracing. *Computer Graphics Forum*, 2021, Vol. 40, No. 2, Pp. 683–712. DOI: 10.1111/cgf.142662

28. **Meister D., Bittner J.** Performance Comparison of Bounding Volume Hierarchies for GPU Ray Tracing. *Journal of Computer Graphics Techniques*, 2022, Vol. 11, No. 3, pp. 1–19.

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Vitaliy G. Chernyi
Чёрный Виталий Григорьевич
E-mail: vitaly.g.cherny@gmail.com
ORCID: https://orcid.org/0009-0003-6281-6726

Marina V. Bolsunovskaya
Болсуновская Марина Владимировна
E-mail: marina.bolsunovskaia@spbpu.com
ORCID: https://orcid.org/0000-0001-6650-6491

Submitted: 22.03.2026; Approved: 17.06.2026; Accepted: 22.06.2026.
Поступила: 22.03.2026; Одобрена: 17.06.2026; Принята: 22.06.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 81–94.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 81–94.

System Analysis and Control
Системный анализ и управление

Research article
DOI: https://doi.org/10.18721/JCSTCS.19208
UDC 658.5:004.9

METHOD FOR IMPLEMENTING SCHEDULING SYSTEMS IN DISCRETE-CONTINUOUS PRODUCTION: THE CASE OF OIL REFINERIES

R.A. Shaidullin ✉

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ renat.a.s@yandex.ru

Abstract. This paper addresses the complex problem of production scheduling for continuous production facilities, such as oil refineries and petrochemical plants. The typical challenges of existing models are analyzed in detail, including the difficulty of accounting for all technological and logistic constraints, operational challenges, and the limited time available for schedule preparation and verification. As a solution, a five-step implementation methodology is proposed: 1) model aggregation to ensure solvability and accuracy, 2) selection of data updating methods (manual, semi-automated, or fully automated), 3) user adaptation through a structured planning process and automatic checks, 4) ensuring additivity between monthly plans based on linear programming and shift-daily schedules, and 5) automatic monitoring of plan and actual asynchrony to detect deviations. The methodology is embedded into the system implementation process, linking methodological principles with each project stage. It constitutes a replicable approach applicable to petrochemical and other continuous production facilities, ensuring the creation of a robust and adequate system capable of operating under the high dynamics and complexity of modern production.

Keywords: production scheduling, oil refinery, planning models, implementation methodology, APS

Citation: Shaidullin R.A. Method for implementing scheduling systems in discrete-continuous production: the case of oil refineries. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 81–94. DOI: 10.18721/JCSTCS.19208

© Shaidullin R.A., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19208
УДК 658.5:004.9

МЕТОД ВНЕДРЕНИЯ СИСТЕМ КАЛЕНДАРНОГО ПЛАНИРОВАНИЯ В ДИСКРЕТНО-НЕПРЕРЫВНЫХ ПРОИЗВОДСТВАХ: ПРИМЕР НПЗ

Р.А. Шайдуллин ✉

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ renat.a.s@yandex.ru

Аннотация. В работе рассматривается комплексная задача календарного планирования непрерывных производств, таких как нефтеперерабатывающие и нефтехимические предприятия. Подробно анализируются характерные проблемы существующих моделей, включая сложность учета всех технологических и логистических ограничений, трудности эксплуатации и ограниченное время на подготовку и верификацию планов. В качестве решения предлагается методика внедрения из пяти шагов: 1) агрегирование модели для обеспечения ее разрешимости и точности, 2) выбор способов актуализации данных (ручной, автоматизированный, автоматический), 3) адаптация пользователей посредством структурированного процесса планирования и автоматических проверок, 4) обеспечение аддитивности между месячными планами на основе линейного программирования и сменно-суточными графиками, 5) автоматический контроль асинхронности плана и факта для выявления отклонений. Методика встроена в процесс внедрения системы, связывая методологические принципы с каждым этапом проекта, и представляет собой тиражируемый подход, применимый к нефтехимическим и другим непрерывным производствам и обеспечивающий создание устойчивой и адекватной системы, способной функционировать в условиях высокой динамики и сложности современного производства.

Ключевые слова: календарное планирование, нефтеперерабатывающий завод, модели планирования, метод внедрения, APS

Для цитирования: Shaidullin R.A. Method for implementing scheduling systems in discrete-continuous production: the case of oil refineries // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 81–94. DOI: 10.18721/JCSTCS.19208

Introduction

The efficiency of continuous production facilities, such as oil refineries and petrochemical plants, depends directly on the consistency of long-term volume planning, operational monthly plans, and shift-level production schedules. Significant financial losses, such as missed delivery deadlines, vessel demurrage, and non-optimal use of storage capacity, are caused by the inability to translate high-level plans into feasible, short-term schedules [1]. Despite the availability of commercial APS/MES solutions, implementations of production scheduling systems remain rare and often do not yield sustainable benefits [2].

The core difficulties are well known: models that fully capture all technological and logistic constraints become computationally intractable; manual updating of large volumes of input data consumes time that is simply not available for short-horizon rescheduling; coordinating models at several planning levels requires substantial manual effort; and plan vs actual analysis cannot be performed with the necessary frequency and depth.

© Шайдуллин Р.А., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

This paper presents a systematic method for implementing production scheduling systems that simultaneously addresses these problems. The method is structured around five interrelated principles:

1) aggregation,
2) actualization,
3) user adaptation,
4) additivity of plans,
5) asynchrony of plan vs actual.

Though the method is described for an oil refinery, it is designed for continuous and semi-continuous process industries in general.

Production scheduling in the process industries has been studied from several directions.

Approaches to the production scheduling problem are divided into deterministic and non-deterministic [3], employing both exact and heuristic solution methods

Mathematical programming

The dominant optimization approach is Mixed-Integer Linear Programming (MILP) [4]. MILP models can explicitly account for unit operating modes, tank capacities, and material quality constraints and have been extended to multi-product inventory logistics [5]. However, full-scale refinery models quickly become computationally intractable, which prevents their use for frequent, short-horizon rescheduling [4].

Simulation

Simulation models reproduce the dynamic behavior of production through rules and state changes [6] and have been employed for operational scheduling and plan vs actual analysis [7]. Among the main approaches to simulation modeling, one can distinguish the object-oriented approach, agent-based modeling, and distributed simulation [8]. The main drawback of simulation is the need for extensive manual control to find an acceptable schedule, making them time-consuming for daily use.

Decomposition and aggregation

To manage complexity, decomposition-coordination methods divide the plant-wide problem into sub-problems by production stages [9]. Aggregation techniques reduce model dimensionality by simplifying resources and time scales [10]. While computationally effective, these approaches introduce the problem of ensuring consistency and feedback between the sub-models and the original problem.

Heuristic methods

Rule-based and metaheuristic algorithms can find acceptable solutions quickly [11], but they sacrifice optimality and require significant tuning by experienced users.

The analysis reveals that existing approaches treat individual aspects of the scheduling problem in isolation – model formulation, algorithm, or data integration. While comprehensive reviews and integrated views on scheduling models and tools exist [12], no holistic implementation methodology has been proposed that simultaneously addresses the choice of model granularity, automation of data actualization, coordination of multiple planning levels, and operational control of plan vs actual deviations. This gap is the starting point of the present work.

Planning models issues

Despite numerous scientific works devoted to production scheduling tasks and the availability of commercial products created to solve these tasks, current implementations of scheduling systems for oil refining productions are rare. Analysis of the features of models and planning systems, data sources, as well as the implementation process at refineries, reveals a number of issues with existing approaches to this task.

Issue 1. The model does not account for all constraints

Sequential Linear Programming (SLP) methods, used in modeling systems such as PIMS by AspenTech[1], RPMS by Honeywell[2], Plan by Nauka LLC[3], and Production Modeling System for Oil Refining and Petrochemistry (SMPR) by Centre of Digital Technologies LLC [13], excellently solve the task of monthly volumetric planning for oil refining production considering flow quality indicators, but are not applicable for the production scheduling task, as the models of these systems do not account for logistical constraints (tank farms, loading racks). Therefore, to account for these constraints, it is necessary to use other models and solution methods.

Issue 2. The model is difficult to use

Simple mathematical planning models may not account for all necessary constraints. For example, using a simple tabular simulation model created in MS Excel, one can easily create a schedule for petroleum product shipments with daily detail, but the capacities of loading racks, the operation of tanks and events lasting less than a day will not be accounted for.

When using complex mathematical models and methods, e.g., optimization MILP models or heuristic methods, problems arise with their configuration to obtain a solution, the input of additional constraints necessary to account for features that emerged after the model was put into operation, as well as the problem of interpreting and analyzing the obtained solution.

After analyzing the obtained solution, a need for its adjustment may arise because either the model does not account for all constraints, or the user possesses knowledge about upcoming events that are unavailable to the model as source data. For example, the priority of producing a certain product, a decrease in unit capacity, or the need to empty a tank below the minimum inventory level. In such a case, the user must either add a constraint to the model or manually correct the obtained solution. When using an optimization model, introducing an additional constraint can lead to a change in the solution over the entire planning horizon or to an infeasible solution. Consequently, the need for re-analysis of the obtained plan or schedule arises. This necessity also arises when using other models and methods.

Issue 3. Limited time for schedule preparation

Planning department employees can spend from one day to a week preparing the monthly volume plan. The preparation process occurs before the start of the planning period, for example, by the 25th of the month preceding the planning period. The monthly production plan undergoes approval and endorsement in several divisions of the plant and the oil company. The calculation of the plan itself using an optimization model takes no more than 5 minutes; however, the preparation of source data, model actualization, analysis of results, and plan coordination take significant time.

Significantly less time is allocated for the preparation of the production schedule. For example, a production schedule with a 24-hour planning period cannot take more than one day to prepare; a schedule with a one-shift planning period cannot take more than one shift; and a plan with a one-hour planning period cannot take more than one hour. Thus, the time for preparing and verifying source data, calculating, and analyzing the obtained results is limited.

When preparing data for calculation using a production scheduling model, data from other models are used, for example, from current planning models and balance models, as well as from various source data systems (Fig. 1).

[1] AspenTech | Process Engineering, Asset Performance Management, Digital Grid Management AI Software | Industrial Digital Transformation, Available: http://aspentech.com/ (Accessed 14.11.2025).

[2] Process Automation | Honeywell, Available: https://process.honeywell.com/ (Accessed 14.11.2025).

[3] IT-kompaniia NAUKA – tsifrovizatsiia protsessov i proizvodstv [IT company NAUKA – digitalization of processes and production], Available: https://ntik.ru/ (Accessed 14.11.2025).

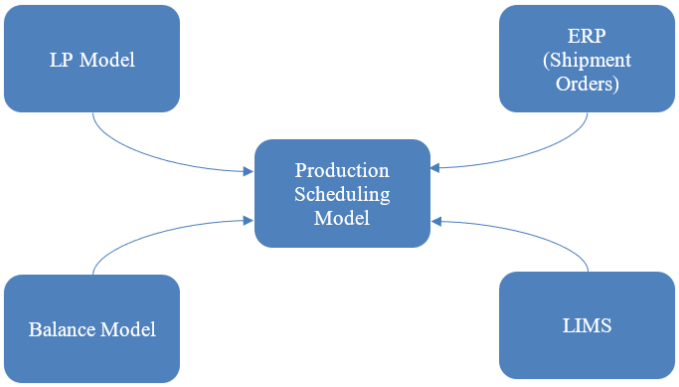


Fig. 1. Interaction of the production scheduling model with other models and data sources

Time during source data preparation can be spent not only on forming the conditions for solving the problem but also on detecting and correcting discrepancies between the production scheduling model and the source models/systems. Thus, a new stream in the LP model, not accounted for in the production scheduling model, can lead to an inadequate production schedule.

When using several models with different time scale, the task of correlating the obtained results for nested time intervals arises. The solution of the monthly planning model does not contain instructions on how and in which period to implement the monthly assignments. Similarly, a production schedule with a daily planning period does not specify how production should function within a day. At the same time, the solution of the monthly planning model contains source data for production scheduling models.

Issue 4. Interrelation of planning models

Solving the production scheduling task as a whole for the entire production with the required detail (in terms of days/shifts, hours/minutes/events) does not seem feasible. For mathematical programming methods, the difficulty lies in the high dimensionality of the models. Simulation modeling methods require a large amount of manual control to find a solution. Heuristic methods do not guarantee achieving an optimal solution. Therefore, the planning task is divided into parts – production stages, for example, feedstock receipt, primary and secondary processing, blending and packaging of finished products, product shipment (Fig. 2).

The process of decomposition into subtasks can be performed not only in terms of parts of the production but also in terms of time (planning horizon, time discretization) and the detail level of production parts (accounting for part of the technological constraints).

When using several models for different parts of the production or with different detail levels according to the technological scheme, it is necessary to consider the interrelation of these models (Fig. 3), and hence the systems ensuring their construction, data population, solution, analysis, and output of results.

Violation of the interrelation of models causes erroneous calculations: operating conditions, as well as the production itself, can change, and therefore the models of this production must change. If model changes occur uncoordinatedly, the solution of the production scheduling task ceases to be adequate.

Issue 5. Plan-vs-actual analysis

The actual operation of production should correspond to the calculations produced using planning models, i.e., the actual should be close to the plan, which is verified by their comparison. When using several models that solve the planning task for different production areas or with different detail, comparing their calculation results with the actual remains relevant but is complicated by the large

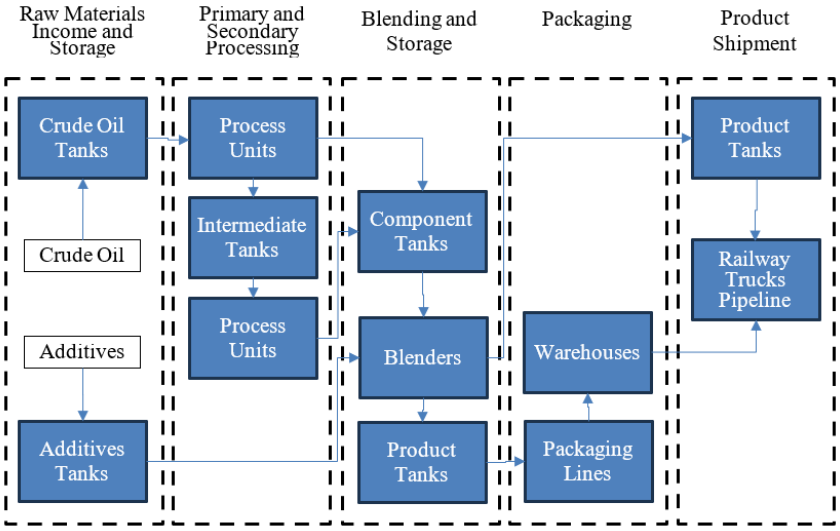


Fig. 2. Example of dividing refinery production into stages

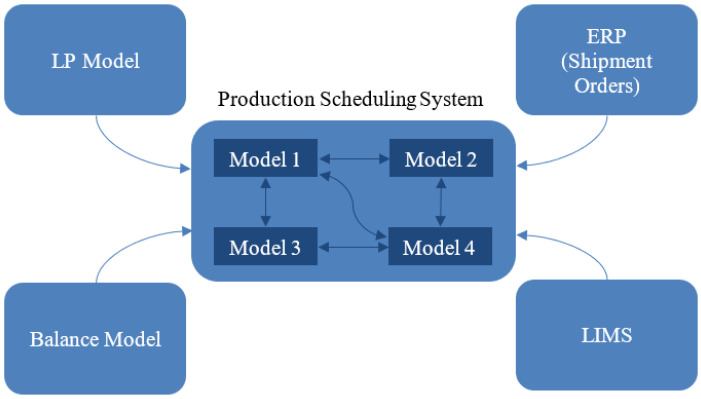


Fig. 3. Interaction of production scheduling models with each other, with other models and data sources

number of compared parameters and the limited time necessary for conducting the plan-vs-actual analysis and making a decision on plan adjustment.

The above list of problems must be taken into account when optimizing the work processes of production scheduling at a refinery, namely when choosing a production scheduling system and the process of its implementation (methodology).

Proposed method

The proposed method consists of five steps, briefly outlined in [14], that must be performed jointly. Each step corresponds to a principle derived from the issues identified in previous section.

Step 1. Model aggregation

A planning model is a description of a technological object, with the help of which one can answer the question of how this object will operate in the future. The model consists of elements of the technological object, such as units, tanks, pipelines, etc. It must account for the parameters of these elements, for example, capacity constraints, volume, flow rate, etc. Furthermore, the model uses for planning a set of source data that establishes the correspondence of the model elements with the

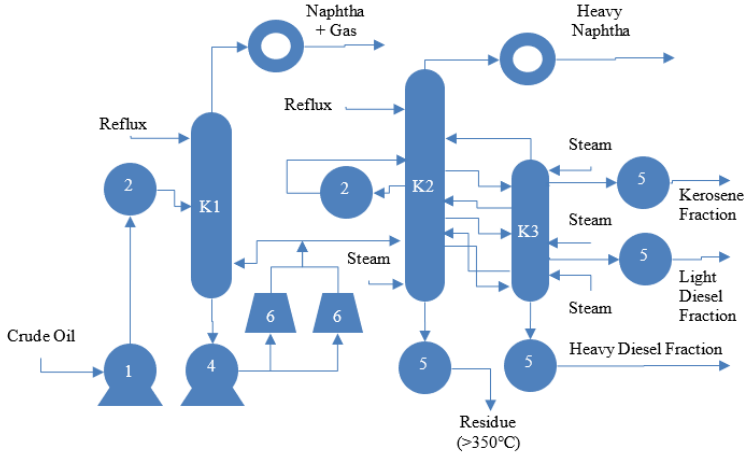


Fig. 4.1. Detailed model of a CDU: 1) crude oil pump; 2) heat exchangers; 3) condensers-coolers; 4) hot pump; 5) chillers; 6) furnaces; K1–3 – distillation columns

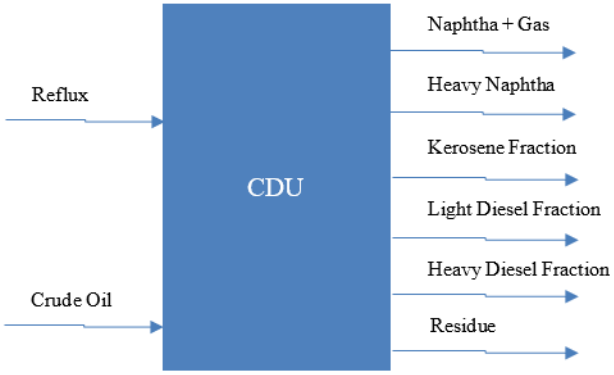


Fig. 4.2. Simplified model of a CDU

current (tank inventories) and target (shipment plan) state of the technological object. Thus, we can speak of three components of the model:

1) element description,
2) parameters,
3) source data.

The same technological object can be described by models of different detail levels.

Firstly, the level of detail is set by the number of elements used to model a particular technological object. For example, an atmospheric-vacuum Crude Distillation Unit (CDU), consisting of pumps, heat exchangers, coolers, furnaces, and columns, can be represented as a single model element with incoming and outgoing petroleum product streams (Fig. 4.1). If the goal is to obtain a plan for this unit's operation requiring forecasting of product output depending on feedstock properties and capacity, there is no need to consider its technological device, and the chosen level of aggregation allows simplifying the model (Fig. 4.2).

Secondly, the number of parameters and constraints related to a single element also determines the model's granularity. To schedule tank farm operations, it is necessary to consider minimum and maximum tank inventory constraints. Additionally, there could be a requirement for the model to consider

the desired minimum and maximum inventory level. Such constraints will increase the complexity of the planning model but also will allow more accurate schedule.

Thirdly, the number of parameters and input data required for a model's validity depends on the number of elements. For example, a model for a 20-tank tank farm requires twice as much data on tank constraints and initial conditions as a 10-tank farm.

Overview of step 1

The level of detail of the scheduling model is chosen based on two criteria: the accuracy required for decision-making and the computational time acceptable for the given rescheduling frequency. Aggregation refers to the number of modeled elements (e.g., representing an entire crude distillation unit as a single node with input/output streams vs detailed flowsheet) and the set of constraints (e.g., hard capacity limits only vs additional desired inventory bands). The chosen level of aggregation directly determines the feasibility of subsequent automatic actualization and user interaction.

Step 2. Data actualization

Actualization means bringing the model into conformity with the current production situation. The model's parts are updated with different frequencies: model elements and their parameters change rather rarely, typically during modernization, repair, or commissioning of new production facilities; source data for calculation, on the contrary, change frequently.

Manual, fully automated, or semi-automated – these are methods of model actualization in consideration.

With the manual method, the user makes the necessary changes to the model using the graphical interface of the planning system. The manual actualization method is characterized by high time consumption, proportional to the model's detail (see aggregation level), as well as a high probability of errors when making changes to the planning model, setting parameters, and entering source data.

The fully automated method involves configuring the information system services so that changes to the planning model are made without user, for example, based on changes in an external application's database or by automatically importing a file with the necessary data. With the fully automated actualization method, it is necessary to establish a relationship between the planning model and data sources and have data integration and validation tools.

The semi-automated method is intermediate between manual and fully automated ones, as it requires the user participation, for example, to select only specific data for import, put additional information to this data, or manually specify the correspondence between the data source and a model element.

An information system that actualizes the model can update the model's components using various methods, presented in Table 1.

Table 1

Example of planning model actualization methods

Model Components	Actualization Method		
	Manual	Fully Automated	Semi-automated
Element's description	✓		
Element's parameters	✓		
Tank Source Data		✓	
Quality Source Data			✓
Order Source Data	✓		

The interrelation of the aggregation level with the choice of actualization method can be demonstrated by the following example. Let a planning model have a parameter specifying the minimum oil product level in a tank. A period in which the oil product level in the tank is below the minimum contains non-zero penalty variables, indicating a violation of this constraint, and the solution is considered inadequate. However, after the completion of scheduled tank maintenance, the inventory in it is less than the constraint set in the model, which leads to the appearance of non-zero penalty values. Eliminating the presence of non-zero penalties in this situation can be achieved in various ways:

1. One can modify the planning model so it will take into account the possibility of violating the minimum tank inventory constraint after coming out of scheduled maintenance. This will affect the model's level of detail and complicate its structure.

2. The planning model remains unchanged, but an actualization method is created where the parameter specifying the minimum tank inventory will be zero or equal to the residue in the tank after the completion of maintenance. In this case, it is necessary to calculate or estimate the time it will take for the tank to fill to the minimum level in order to restore the real constraint value.

The second option can be implemented with fully automated and semi-automated methods of actualizing source data and model parameters. With the manual actualization method, this option would require significant time from the model user and could be a source of model configuration errors if the user does not return the constraint value to its original state.

When choosing a method for actualizing source data, the frequency of calculation using the model is decisive: when calculating once a month/week, there is enough time for the manual actualization method; when calculating once a day/hour and with a large volume of data, there will not be enough time for manual data update, validation, and entry into the model. Thus, to use the manual method of source data actualization, it is necessary either to reduce their volume, i.e., aggregate the model (see aggregation level), which will entail a decrease in the adequacy of the obtained plans, or to perform calculations less frequently (once a week), which will entail an increase in reaction time to changes in the production situation and reduce planning efficiency. Fully automated and semi-automated actualization methods solve these problems; however, they require establishing relationships between data sources and the planning model and using integration tools.

Overview of step 2

For each component of the model (structure, parameters, initial data etc.) a concrete method of update (manual, fully automated, and semi-automated) is fixed according to the frequency of change and the available integration infrastructure. For example, laboratory quality data are updated automatically from LIMS, while the activation of a new process unit is entered manually. The decision is critical: with shift- or hour-level rescheduling, manual data entry becomes impossible without sacrificing timing or accuracy. Fully automated and semi-automated actualization eliminates this bottleneck and prevents configuration errors caused by missed or delayed updates.

Step 3. User adaptation

As with the implementation of any complex information system, users are main reasons of the success or failure of the project. The adoption of the system, as well as new workflows, also depends on the system's users.

On the one hand, users can undergo formal training on the use of the system and new workflows, in the form of a user course, and gain an understanding of how to use the scheduling system for their daily tasks. However, it is worth noting that although this approach will allow the system to be handed over for industrial operation and used successfully for some time after commissioning, it does not guarantee the maintenance of a lasting effect from the implementation of the new information system.

On the other hand, the participation of key end users in the project implementation process, undergoing in-depth training in the form of a series of advanced user courses, will allow them to view the system not as a "black box", but as a flexible tool for performing and optimizing daily work tasks.

At the same time, the topics of advanced training must include model creation, mathematical aspects of finding a solution, configuration, and system support. This approach requires a certain level of training from the users.

Regarding the optimization of business processes, in addition to the requirements for the level of training of end users at the time of handing over the system for trial operation, it is also necessary to consider the degradation of user knowledge over time, or the complete disappearance of this knowledge when team members change, which happens quite often due to promotions of planning department employees or staff turnover.

Furthermore, training materials, courses, and user support materials must be updated with product updates, and this imposes certain requirements on the software vendor.

The level of detail of the planning workflow, which is the result of optimization/re-engineering of scheduling business processes, is directly related to the user and their level of mastery of the system. The planning workflow must be sufficiently detailed, but not overloaded, so that it can be comprehended. The aggregation level and the model actualization method directly influence the level of detail of the planning workflow, which, in turn, affects the efficiency of using the implemented planning system.

Overview of step 3

The planning workflow and system interface are designed so that the user perceives the system as a transparent decision-support instrument, not as a "black box". The workflow must be sufficiently detailed to reflect the actual work of the planning department but not overloaded. Special attention is paid to the degradation of user knowledge over time and during staff rotation; therefore, the workflow is reinforced by automated validation checks that prevent typical human errors.

Step 4. Additivity of plans

Additivity is a property of quantities, consisting in the fact that the value of a quantity corresponding to a whole object is equal to the sum of the values of the quantities corresponding to its parts.

Planning is always conducted at several levels. Starting from investment plans for several years and ending with hourly schedules or event-based periods. The planning workflow, which is the result of optimization/re-engineering of planning business processes, is responsible for the sequence of actions during planning, the identification of stakeholders, and data flows in planning.

The planning workflow must consider the sequence of plan calculations, and this means the interaction of several plans, when the results of one are the source data for the calculation of another, and then the results of the second adjust the first, thus establishing feedback. The interaction of plans is the interaction of several systems (and/or their models) of different planning levels, or of one model, but calculated for different source data (which is quite rare). At the same time, the workflow must consider the detail level of the models, the method of actualization, and user adaptation.

Overview of step 4

Consistency between the results of models at different planning levels is enforced procedurally and mathematically. The monthly LP plan serves as a set of target volumes for the scheduling model, while the schedule, once calculated, provides feedback on the achievability of those targets. This creates a closed-loop management cycle and eliminates the situation where the schedule contradicts the monthly plan.

Step 5. Asynchrony of plan and actual

Planning is a corrective process, so it is necessary to know the current state, the target state, the deviation, and, moreover, to respond appropriately to identified deviations. The planning system must automatically account for the plan deviation from the actual, considering the actualization method: perform data collection, comparison, analysis, and presentation of results in the form of reports or recommendations. Report generation must satisfy the principle of user adaptation, and recommendations for adjustment must satisfy the principle of plan additivity. In addition, actual data on the

production state may come from systems that have a different object detail, which establishes a connection with the aggregation principle.

Overview of step 5

A mechanism for the automatic detection of deviations is built into the system. Scheduled and actual inventories, material movements, and shipments are compared at every calculation cycle. Discrepancies are visualized, and rule-based recommendations for plan adjustments are generated. This automates the labor-intensive plan-vs-actual analysis and guarantees that the schedule remains relevant under changing production conditions.

Advantages of the proposed method

The five-step method provides several measurable and qualitative advantages over conventional, isolated implementation approaches. A comparative summary is given in Table 2.

Table 2

Comparison of approaches

Criterion	MILP	Standalone Simulation	Decomposition	Proposed Method
Solvability for large problems	Low	High (manual search)	Medium	High (guided aggregation)
Schedule preparation time	Long	Long	Medium	Short (automated actualization)
User expertise required	Very high	Medium	High	Medium (structured workflow)
Cross-model consistency	Low	Low	Medium	High (additivity principle)
Plan-vs-actual analysis	Manual	Semi-automated	Manual	Automated (asynchrony)

Quantitatively, pilot applications at a refinery site showed a reduction in schedule preparation time from approximately eight hours to three to four hours, and the elimination of data transfer errors between the monthly plan and the daily schedule. The key qualitative advantage is the creation of a single, consistent version of the production plan across all management levels, which is automatically maintained over time through the principles of actualization and asynchrony.

Implementation framework

The implementation of a production scheduling system is carried out similarly to the implementation of any information system for production automation, and both agile methods, the waterfall approach, and a hybrid approach can be applied [15]. The stage-wise sequence presented below corresponds to the classic process of creating automated systems, regulated by GOST 34.601-90 “Automated systems. Stages of development” or ISO/IEC 12207 “Systems and software engineering – Software life cycle processes”. The project stages when implementing the waterfall and hybrid approach are the following:

1. Site survey and technical requirements development.
2. System design.
3. Development and actualization of production scheduling models.
4. Model testing of the system.
5. System deployment on site.
6. System integration.
7. User training.
8. Trial operation.
9. Model adjustment and handover of the system for industrial operation.

When implementing a hybrid approach, stages 1–9 can be carried out iteratively and only for a specific functionality of the system being developed in a sprint.

At the first stage, a survey of the refinery is carried out, during which the production technology, available documentation, technological parameters, statistical data on equipment loading, production volumes, feedstock supply, and product shipment are studied. Data on laboratory operations are also collected. Information is gathered about the plant's information systems that are sources and consumers of data for the planning system. Also, during the first stage, the technical requirements for the system are developed and approved, detailed based on the survey results.

At the second stage, system design is carried out, during which a technical (or preliminary) design is developed. The technical design includes the system architecture, component interaction diagrams, integration interfaces with other systems, and specifications of equipment and software. The preliminary design contains preliminary solutions and concepts that are detailed in the technical design. At this stage, the planning workflow is also developed.

At the third stage, based on the approved design solutions, the development and actualization of production scheduling models are carried out. The models are populated with data, parameterized, and configured in accordance with the plant's technological processes.

At the fourth stage, model testing of the system is conducted. Testing is performed on an isolated stand using actual and planned data to verify the correctness of calculations, the adequacy of the models, and the functionality of the planning algorithms.

At the fifth stage, the system is deployed at the plant site. Server and client software are installed and configured, databases are deployed, and the infrastructure is prepared.

At the sixth stage, work is carried out to integrate the production scheduling system with other enterprise systems (e.g., ERP, MES, LIMS), which are sources and consumers of data. Automated information exchange flows are configured.

At the seventh stage, users are trained to work with the system, configure it, update existing planning models, and create new models.

At the eighth stage, the system is handed over for trial operation. At the plant, the adequacy of the production scheduling models under real conditions is checked, software and model errors are identified, and requirements for additional reports and functionality are formed.

At the ninth stage, based on the results of trial operation, the system's models and software are adjusted, necessary reports are created, and documentation, including the planning workflow, is updated. After eliminating all identified problems and successfully completing acceptance tests, the system is handed over for industrial operation.

The nine-stage implementation process is not a mere project plan but the operationalization of the five-step method. The Aggregation level is formalized during Site Survey and Design. The Actualization method is configured during Integration. User Adaptation is realized through training and trial operation. Additivity and Asynchrony are embedded as functional requirements during System Design and are validated during Model Testing and Trial Operation. This explicit link between the methodological principles and the project stages distinguishes the approach from conventional system deployments and is the critical factor for long-term sustainability.

Conclusion

The analysis presented in this paper revealed that existing approaches to production scheduling fail in industrial practice not because of model inaccuracy, but because of the absence of a holistic implementation methodology that simultaneously addresses model aggregation, data actualization, plan consistency, and user involvement.

The main result of the work is a novel five-step implementation method (aggregation, actualization, user adaptation, additivity, asynchrony) that systematically guides the design, deployment, and

operation of a scheduling system. Unlike previous works that focus on isolated algorithmic performance, this method ensures the system's viability in a real production environment.

Application of the method led to a reduction of schedule preparation time by approximately 40%, complete elimination of errors caused by manual data transfer between planning levels, and the creation of a single, synchronized planning cycle. The method provides a replicable framework that is currently being extended to other continuous process industries, confirming its generality beyond oil refining.

REFERENCES

1. **Petukhov M.Yu.** On the Relevance of Implementation of the Automated System of Production and Supply Sheduling at Oil Refinery. *Automation and Remote Control*, 2018, Vol. 79, Pp. 1168–1174. DOI: 10.1134/S0005117918060140

2. **Shobrys D.E., White D.C.** Planning, scheduling, and control systems: Why can they not work together. *Computers & Chemical Engineering*, 2000, Vol. 24, No. 2–7, Pp. 163–173. DOI: 10.1016/S0098-1354(00)00508-1

3. **Lei D.** Multi-objective production scheduling: A survey. *The International Journal of Advanced Manufacturing Technology*, 2009, Vol. 43, Pp. 926–938. DOI: 10.1007/s00170-008-1770-4

4. **Floudas C.A., Lin X.** Mixed Integer Linear Programming in Process Scheduling: Modeling, Algorithms, and Applications. *Annals of Operations Research*, 2005, Vol. 139, Pp. 131–162. DOI: 10.1007/s10479-005-3446-x

5. **Zyngier D., Kelly J.D.** Multi-Product Inventory Logistics Modeling in the Process Industries. In: *Optimization and Logistics Challenges in the Enterprise* (eds. W. Chaovalitwongse, K.C. Furman, P.M. Pardalos), 2009, Vol. 30, Pp. 61–95. DOI: 10.1007/978-0-387-88617-6_2

6. **Gurko A.V., Sotnikov V.V., Lisitsyn N.V.** Imitatsionnoe modelirovanie operativnogo upravleniia proizvodstvom [Simulation modeling of operational production management]. *Izvestiia Orlovskogo gosudarstvennogo tekhnicheskogo universiteta. Seriia: Informatsionnye sistemy i tekhnologii* [*Information Systems and Technologies]*. 2006. Vol. 1–2, Pp. 37–46.

7. **Shaidullin R.A., Khokhlov A.S., Prokazina M.V.** Simulation models in oil refinery scheduling system. *Automation in Industry*, 2012, Vol. 10, Pp. 15–21.

8. **Pogorelov A.S.** Imitatsionnoe modelirovanie: osnovnye podkhody i instrumental'nye sredstva [Simulation modeling: main approaches and tools]. *Izvestiia Orlovskogo gosudarstvennogo tekhnicheskogo universiteta. Seriia: Informatsionnye sistemy i tekhnologii* [*Information Systems and Technologies*], 2006, Vol. 1, No. 4, Pp. 168–173.

9. **Aonuma T., Nishi T., Takai E.** An implementation of a decomposition-coordination approach to refinery production scheduling. *International Journal of Production Research*, 1983, Vol. 21, No. 6, Pp. 845–857. DOI: 10.1080/00207548308942417

10. **Váncza J., Kis T., Kovács A.** Aggregation – the key to integrating production planning and scheduling. *CIRP Annals*, 2006, Vol. 53, No. 1, Pp. 377–380. DOI: 10.1016/S0007-8506(07)60720-3

11. **Antamoshkina E.A., Sharypova K.V.** The heuristic algorithm for discrete production scheduling. *Information Science and Control Systems*, 2010, Vol. 23, No. 1, Pp. 67–73.

12. **Framinan J.M., Leisten R., Ruiz García R.** *Manufacturing Scheduling Systems: An integrated view on models, methods, and tools*. London: Springer, 2014. DOI: 10.1007/978-1-4471-6272-8

13. **Khokhlov A.S., Mishutin D.Yu., Baulin E.S.** Realizatsiia sistem klassa APS dlia NPZ/NKHK i ikh ob"edinenii [Implementation of APS class systems for oil refineries/petrochemical complexes and their associations]. *Upravlenie razvitiem krupnomasshtabnykh sistem* [*Managing the development of large-scale systems*] (*MLSD'2023*): *Proceedings of the 16th International conference*, 2023, Pp. 94–103. DOI: 10.25728/mlsd.2023.0094

14. **Khokhlov A.S., Konnov A.I., Shaidullin R.A.** Integrated approach to continuous production planning. *Automation in Industry*, 2015, Vol. 4, Pp. 36–40.

15. **Ermakov E.** Application of hybrid methodologies in project management. *Universum: Economics and Law*, 2024, Vol. 119, No. 9, Pp. 15–20. DOI: 10.32743/UniLaw.2024.119.9.18167

INFORMATION ABOUT AUTHOR / СВЕДЕНИЯ ОБ АВТОРЕ

Renat A. Shaidullin
Шайдуллин Ренат Анварович
E-mail: renat.a.s@yandex.ru

Submitted: 20.02.2026; Approved: 20.05.2026; Accepted: 01.06.2026.
Поступила: 20.02.2026; Одобрена: 20.05.2026; Принята: 01.06.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 95–108.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 95–108.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19209
UDC 004.891.3:519.853

STRUCTURE-AWARE DECISION SUPPORT SYSTEM FOR MOO-DRIVEN CONTROL

A. Deeb ✉ , *V.N. Khokhlovskiy*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ dib_a@spbstu.ru

Abstract. Cyber-physical control systems must satisfy strict constrains on the magnitude of control actions and the rate of signal change, while simultaneously balancing multiple, often conflicting, control objectives. In multi-objective optimization (MOO) problems, the optimizer naturally produces a set of competing control candidates. However, in closed-loop operation, a single implementable control action must be selected at each sampling instant. In practice, this selection is typically performed through fixed scalarization rules, which can lead to performance degradation under changing operating conditions in the case of non-convex Pareto fronts or poorly scalable objectives. This work aims to improve the quality of control based on MOO through a decision support system that accounts for the structure: the proposed architecture decouples online Pareto-front construction from action selection. A heuristic multi-objective optimization algorithm based on the NSGA-II genetic algorithm generates feasible control sequences under magnitude and rate constraints, while a learned selector, denoted as Learned Pareto Filter (LPF), evaluates these candidates using structural features that account for both individual characteristics of the solutions and their interrelations within the Pareto set. The method is evaluated on two benchmark problems. First, a system with an analytically known Pareto front structure is used. Second, the approach is validated on a nonlinear boiler-turbine load-changing scenario, widely used in constrained nonlinear control studies. The results demonstrate that LPF-based selection produces smoother control actions, improved tracking performance, and greater robustness to regime changes compared to objective-based scalarization methods. The findings confirm that, in Pareto-based control, closed-loop quality depends critically on the selection mechanism rather than solely on the optimizer generating the Pareto set.

Keywords: multi-objective optimization, model predictive control, decision support systems, Pareto selection, graph neural networks, constrained nonlinear control

Acknowledgements: The research was financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the program “World-class Research Center: Advanced Digital Technologies” (Agreement No. 075-15-2022-311 dated 20.04.2022).

Citation: Deeb D.A., Khokhlovskiy V.N. Structure-aware decision support system for MOO-driven control. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 95–108. DOI: 10.18721/JCSTCS.19209

© Deeb D.A., Khokhlovskiy V.N., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19209
УДК 004.891.3:519.853

СТРУКТУРНО-ОРИЕНТИРОВАННАЯ СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ДЛЯ УПРАВЛЕНИЯ, ОСНОВАННОГО НА МНОГОКРИТЕРИАЛЬНОЙ ОПТИМИЗАЦИИ

А. Диб ✉ , *В.Н. Хохловский*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ dib_a@spbstu.ru

Аннотация. Киберфизические системы управления должны удовлетворять строгим ограничениям по величине управляющих воздействий и скорости изменения сигнала, одновременно обеспечивая баланс между несколькими (часто противоречивыми) целями управления. В задачах многокритериальной оптимизации оптимизатор естественным образом генерирует множество конкурирующих кандидатов на управление. Однако в замкнутом контуре на каждом шаге дискретизации требуется выбрать единственное реализуемое управляющее воздействие. На практике этот выбор обычно выполняется с помощью фиксированных правил скаляризации, что может приводить к ухудшению характеристик при изменении режимов эксплуатации в случае невыпуклой формы фронта Парето или трудно масштабируемых критериях. Данная работа направлена на повышение качества управления, основанного на многокритериальной оптимизации, на базе системы поддержки принятия решений, учитывающей структуру: предложенный метод разделяет этап онлайн-построения фронта Парето и этап выбора управляющего воздействия. Эвристический алгоритм многокритериальной оптимизации, основанный на генетических алгоритмах NSGA-II, генерирует допустимые последовательности управляющего воздействия с учетом ограничений по величине и скорости изменений, тогда как предложенный обучаемый селектор, обозначенный как Learned Pareto Filter (LPF), оценивает этих кандидатов, используя структурные признаки, учитывающие как индивидуальные характеристики решений, так и их взаимосвязи внутри множества Парето. Метод оценивается на двух эталонных задачах. Во-первых, используется система с аналитически известной структурой фронта Парето. Во-вторых, подход проверяется на нелинейной задаче изменения нагрузки котлотурбинной установки, широко используемой в исследованиях управления нелинейными системами с ограничениями. Результаты показывают, что выбор на основе LPF обеспечивает более гладкие управляющие воздействия, улучшенную точность слежения и большую устойчивость к изменениям режима по сравнению с методами, основанными исключительно на скаляризации критериев. Полученные выводы подтверждают, что качество Парето-ориентированного управления замкнутым контуром в критической степени определяется механизмом выбора, а не только оптимизатором, формирующим множество Парето.

Ключевые слова: многокритериальная оптимизация, модельно-прогностическое управление, системы поддержки принятия решений, отбор по Парето, графовые нейронные сети, управление нелинейными системами с ограничениями

Финансирование: Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках программы «Научный центр мирового уровня: Перспективные цифровые технологии» (договор № 075-15-2022-311 от 20.04.2022).

Для цитирования: Deeb D.A., Khokhlovskiy V.N. Structure-aware decision support system for MOO-driven control // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 95–108. DOI: 10.18721/JCSTCS.19209

© Диб А., Хохловский В.Н., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

Introduction

New-generation cyber-physical systems are increasingly required to operate within strict safety envelopes, under hard actuator bounds and rate constraints, while simultaneously satisfying multiple, and often conflicting, performance objectives. This requirement is particularly pronounced in energy and process industries, where strong variable coupling, nonlinear dynamics, and continuously shifting operating regimes render controller design a nontrivial task. Thermal power units and boiler-turbine systems constitute representative examples: they must guarantee safety-critical constraints while delivering the demanded power output despite nonlinear interactions and multivariable dependencies [1–3].

Such characteristics naturally motivate control paradigms that explicitly incorporate constraints and performance trade-offs, rather than reducing the problem to a single aggregated objective. In this context, Model Predictive Control (MPC) has emerged as a principled framework for constraint-aware decision making and has been widely adopted in industrial practice [4, 5]. Nevertheless, in highly nonlinear plants, real-time feasibility often necessitates structural simplifications, including local linearization, gain scheduling, or fuzzy approximations around operating points [6–9]. Alternative approaches based on learning-based controllers and evolutionary optimization offer improved adaptability to nonlinearities and uncertainties, yet they frequently introduce considerable computational overhead, which complicates their deployment in strict real-time environments [10–15]. Consequently, practical control design becomes a balance between model expressiveness (nonlinear representations, multi-objective formulations, hard constraints) and the need for immediate actionable decisions.

A common strategy for handling multiple objectives is scalarization, where several performance criteria are combined into a single cost function via weighted sums or related aggregation rules. Although convenient, scalarization requires the selection of preference weights *a priori* and may become fragile when the appropriate trade-off varies across operating regimes [16]. In online control settings, this leads to what may be termed a meta-decision gap: even if a multi-objective optimizer produces a set of nondominated solutions, the closed-loop implementation requires selecting exactly one control input at each sampling instant. If the selection mechanism does not reflect the current regime, the resulting closed-loop behavior may degrade, even when the computed Pareto set is accurate.

To address this meta-decision gap, this paper proposes a structure-aware Decision Support System (DSS) for multi-objective optimization (MOO)-driven control. The proposed DSS operates within a receding-horizon framework. At each step, it constructs an approximate Pareto front of feasible control sequences that satisfy hard input bounds and slew-rate constraints, and subsequently selects one candidate using a learned, structure-exploiting selection mechanism. Instead of relying on fixed scalarization, Pareto selection is formulated as a context-dependent decision problem over a variable-sized set of alternatives. The selector leverages both per-candidate attributes and relational information among candidates – such as dominance relations, smoothness properties, and similarity measures – to adaptively determine the applied control action.

The main contributions of this work can be summarized as follows:

- A set-based formulation for online MOO-driven control under hard input bounds and rate constraints, where the optimizer returns a feasible Pareto set of control sequences and a dedicated selector determines the applied input.
- A structure-aware selection mechanism that evaluates Pareto candidates using features capturing both individual and relational properties, enabling adaptive and regime-dependent trade-off selection without manual weight tuning.
- An experimental validation based on an online Pareto-front construction procedure (NSGA-II) [17], with comparisons against standard scalarization-based selection rules, demonstrating improved decision consistency under changing operating conditions.

Related works

Constraint handling is central in real systems due to actuator saturation and rate limits. MPC provides a principled framework to embed such constraints directly in control synthesis [4–5]. In boiler-turbine and related thermal systems, constrained predictive control has often relied on model simplifications, including fuzzy multi-model schemes and operating-point strategies [3, 6–7], as well as robust and nonlinear variants built on local models or auxiliary constructions to preserve feasibility and stability [18–19]. Online linearization and piecewise representations remain common tools to balance accuracy and computational tractability across regimes [8–9].

When objectives are nonconvex, discontinuous, or expensive to evaluate, evolutionary computation offers flexible black-box optimization and has been applied to control design and online tuning [13–14]. Learning-based controllers can further capture nonlinearities and uncertainties, improving performance over classical baselines [10–11, 20]. However, real-time deployment is constrained by training, inference, and repeated optimization costs within the loop [12], motivating architectures that separate expensive policy search from fast online execution while preserving constraint awareness and regime adaptability.

MOO naturally arises in constrained control problems involving competing criteria such as tracking, energy usage, and safety margins [16]. Evolutionary multi-objective algorithms are widely used to approximate Pareto fronts in complex spaces [17], yet they do not resolve the core online requirement: selecting a single admissible action from a set of nondominated candidates. In practice, scalarization is the dominant baseline, but it encodes fixed global preferences, can miss nonconvex regions, and is sensitive to scaling and regime changes [16, 21].

Theoretical works on multi-objective MPC analyze how selection rules affect stability and performance. Formal multi-objective MPC formulations link Pareto concepts with receding-horizon control [22], utopia-based selection has been used to reason about stability [23], and hierarchical/lexicographic prioritization has been studied to avoid destabilizing trade-offs in nonlinear MPC [24]. Moreover, it has been shown that closed-loop guarantees may be achieved without approximating the entire Pareto front at each step, provided that an appropriate Pareto-optimal sequence is selected [25]. This highlights the importance of the selection mechanism itself.

From a decision-support and multi-criteria decision-making (MCDM) perspective, choosing among Pareto candidates depends on scaling, geometry, and operating context, and fixed scalarizations are often fragile in practice [21]. Recent learning-based approaches, including multi-objective reinforcement learning (MORL), treat Pareto coverage and online selection as primary objectives, guiding policy search with predictive signals to improve front quality and adaptability [26]. However, such methods typically do not emphasize explicit constraint feasibility and receding-horizon structure required in hard real-time systems.

Taken together, the literature reveals a persistent gap: classical multi-objective MPC provides constraint-aware predictive structure and stability analysis under fixed selection principles [21–25], while learning-based multi-objective control emphasizes adaptive selection but often abstracts away strict constraint handling [26]. A structure-aware decision support system for MOO-driven control lies between these paradigms: it preserves explicit constraint feasibility and short-horizon predictive optimization, yet replaces fragile fixed scalarizations with learned, context-aware selectors that operate directly on the structure of the Pareto candidate set.

Problem formulation

Consider a discrete-time controlled system:

$$x_{k+1} = f\left(x_k, u_k\right), \quad x_k \in \mathbb{R}^n, \; u_k \in \mathbb{R}^m, \tag{1}$$

operating under hard input bounds and rate (slew) limits:

$$u_{\min} \le u_k \le u_{\max}, \quad \Delta u_{\min} \le u_k - u_{k-1} \le \Delta u_{\max}, \tag{2}$$

where inequalities are elementwise. At each time step, the controller must choose an admissible action that balances multiple, potentially conflicting objectives. Let

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_p(x) \end{bmatrix} \in \mathbb{R}^p \tag{3}$$

denote a vector of performance criteria to be minimized. For a horizon H, we define a candidate control sequence:

$$U = (u_0,\ u_1, \ldots,\ u_{H-1}) \in \mathbb{R}^{Hm}, \tag{4}$$

and its predicted terminal state:

$$x_H = \Phi_H(x_k, U); \quad \Phi_H(x_k, U) := f(\cdot, u_{H-1}) \circ \cdots \circ f(\cdot, u_0)(x_k). \tag{5}$$

The induced multi-objective cost of the candidate is then:

$$J(U; x_k) := g(x_H). \tag{6}$$

To reduce actuator chattering and encourage smooth actions we add a quadratic increment penalty

$$S(U; u_{k-1}) := \sum_{i=0}^{H-1} \|u_i - u_{i-1}\|_2^2, \quad u_{-1} := u_{k-1}. \tag{7}$$

The online decision problem at time k can be stated as a constrained MOO over sequences:

$$\min_U \left(J(U; x_k), \lambda_s S(U; u_{k-1}) \right), \tag{8}$$

where U satisfies (2) across the horizon, and $\lambda_s \ge 0$ trades off smoothness.

The meta-decision gap in Pareto-based control

A central issue in Pareto-based control is that the optimizer returns alternatives, but the closed loop requires one actionable control input. In many Multiple-Input and Multiple-Output (MIMO) applications, it is common to collapse the objective vector into a single scalar cost via a fixed linear combination (weighted sum),

$$\tilde{J}(U_k) = \sum_{j=1}^{q} w_j J^{(j)}(U_k; z_k), \quad w_j \ge 0. \tag{9}$$

While convenient, such scalarizations implicitly assume that preferences are known *a priori* and remain valid across operating regimes. Moreover, linear scalarizations can fail to represent important trade-offs and often require extensive retuning when conditions change. In practice, this “merge-the-

objectives" step becomes a bottleneck: even if $\mathcal{P}_k$ is accurate, a poor selection rule can degrade closed-loop performance.

Learning-based preference modeling

The proposed strategy is to learn a selector π_0 offline from preference labels generated by an oracle utility $u^*(\cdot)$:

$$u_k^*(i) = u^*\left(c_k^{(i)}; s_k\right), \tag{10}$$

designed to reflect tolerance-aware preferences. A function approximator is trained to predict $\hat{u}_k(i)$ from $(\mathcal{G}_k, i)$, and the online decision is

$$i^* = \arg\max_{i \in \mathcal{V}_k} \hat{u}_k(i), \quad u_k = u_{k|k}^{\left(i^*\right)}. \tag{11}$$

This decoupled architecture replaces fixed scalarization by adaptive, data-driven selection while preserving the underlying multi-objective optimizer. Rather than collapsing to a single scalar objective. We approximate the Pareto set of nondominated candidates at each step. Specifically, we define Pareto dominance in objective space: for two candidates U^a and U^b,

$$\begin{aligned} U^a \prec U^b \Leftrightarrow J\left(U^a; x_k\right) \le J\left(U^b; x_k\right) \ \text{ elementwise and} \\ \exists j : J_j\left(U^a; x_k\right) < J_j\left(U^b; x_k\right). \end{aligned} \tag{12}$$

The (approximate) Pareto front at time k is then the set of candidates not dominated by any other candidate in the population:

$$\mathcal{P}(x_k) := \left\{U \in \mathcal{U}_H(x_k) : \nexists \tilde{U} \in \mathcal{U}_H(x_k) \text{ s.t. } \tilde{U} \prec U\right\}, \tag{13}$$

where $\mathcal{U}_H(x_k)$ denotes the set of feasible sequences under (2). Given $\mathcal{P}(x_k)$, a selector policy must choose one candidate to apply in receding-horizon fashion:

$$U_k^* = \pi\left(\mathcal{P}(x_k), x_k, u_{k-1}\right), \quad u_k = \left(U_k^*\right)_0. \tag{14}$$

Graph neural networks as Pareto-front selectors

A key difficulty in (14) is that the Pareto front is a set whose size and composition vary over time. Hand-designed scalarizations such as weighted-sum and Chebyshev methods impose a fixed preference structure and can behave poorly when the front is nonconvex, poorly scaled, or when the appropriate trade-off changes with the state.

Graph neural networks (GNNs) offer a natural way to map a variable-size set of candidates to scores while exploiting relational structure. A GNN operates on a graph $G = (V, E)$ with node features $\{h_v\}_{v \in V}$ and edge features $\{e_{uv}\}_{(u, v) \in E}$, and updates node embeddings through message passing:

$$m_v^{(\ell)} = \sum_{u \in \mathcal{N}(v)} \psi^{(\ell)}\left(h_v^{(\ell)}, h_u^{(\ell)}, e_{uv}\right); \tag{15}$$

$$h_v^{(\ell+1)} \& = \phi^{(\ell)}\left(h_v^{(\ell)}, m_v^{(\ell)}\right), \tag{16}$$

for layers $\ell = 0,\ldots, L-1$, where $\mathcal{N}(v)$ denotes neighbors of v. A permutation-invariant readout then produces a graph-level embedding

$$z(G) = \rho\left(\left\{h_v^{(L)}\right\}_{v\in V}\right), \tag{17}$$

and finally, a scalar score $s(G)$ is computed by a small predictor head. This architecture is well-suited to Pareto selection because:

1) the candidate set is naturally represented as nodes,
2) pairwise relations can be encoded as edges,
3) invariance to candidate ordering is required.

LPF filtering under feasibility constraints

We propose a learned selector, denoted as Learned Pareto Filter (LPF), that chooses u_k by scoring Pareto candidates using a GNN. At time k, a multi-objective evolutionary search (NSGA-II) produces an approximate front

$$\hat{\mathcal{P}}(x_k) = \left\{U^{(1)},\ldots, U^{(N_k)}\right\}, \tag{18}$$

with associated per-candidate attributes

$$\left(x_H^{(i)},\ J^{(i)} = g\left(x_H^{(i)}\right),\ S^{(i)} = S\left(U^{(i)}; u_{k-1}\right)\right). \tag{19}$$

We build a graph G_k whose nodes correspond to candidates $U^{(i)}$. Node features encode both candidate-specific and context features:

$$h_i = \varphi\left(x_k,\ u_{k-1},\ x_H^{(i)},\ J^{(i)},\ S^{(i)}\right), \tag{20}$$

and edges encode candidate-to-candidate relations, such as dominance indicators, objective-space distances, or normalized differences in terminal states:

$$e_{ij} = \eta\left(J^{(i)} - J^{(j)},\ x_H^{(i)} - x_H^{(j)}\right). \tag{21}$$

The LPF model produces a scalar score for each node i:

$$s_i = GNN_\theta\left(G_k; i\right), \tag{22}$$

and selects the candidate with minimal score:

$$i^* = \arg\min_{i\in\{1,\ldots,N_k\}} s_i,\ u_k = \left(U^{(i^*)}\right)_0. \tag{23}$$

Training signal

To ground the learned selector with a concrete notion of progress, we train θ to imitate a teacher that is available in simulation. For the benchmark considered in this work, the teacher evaluates a candidate by the one-step improvement toward the true Pareto set in decision space, optionally with a smoothness term:

$$y_k^{(i)} = d_{\mathcal{P}}\left(x_{k+1}^{(i)}\right) + \alpha\left\|u_0^{(i)} - u_{k-1}\right\|_2^2,\ x_{k+1}^{(i)} = f\left(x_k, u_0^{(i)}\right), \tag{24}$$

where $d_{\mathcal{P}}(\cdot)$ is a known distance-to-Pareto-set metric in the benchmark and $\alpha \geq 0$ controls smoothness. The training objective is mean-squared error

$$\min_{\theta} \mathbb{E}\left[\left(GNN_{\theta}\left(G_k; i\right) - y_k^{(i)}\right)^2\right], \tag{25}$$

evaluated over graphs produced by rollouts in which the teacher closes the loop. At test time, only $\hat{\mathcal{P}}\left(x_k\right)$ and the learned model are used.

Experiments

Decision support effect towards a known Pareto front

To evaluate the decision-support capability of the proposed selector independently of plant complexity, we adopt a controlled benchmark with analytically known Pareto structure. The system is a discrete-time integrator:

$$x_{k+1} = x_k + T_s u_k, \quad x_k \in \mathbb{R}^2, \quad u_k \in \mathbb{R}^2, \tag{26}$$

subject to elementwise magnitude and slew-rate constraints

$$u_{\min} \leq u_k \leq u_{\max}, \quad \Delta u_{\min} \leq u_k - u_{k-1} \leq \Delta u_{\max}. \tag{27}$$

Two conflicting quadratic objectives are defined at the terminal predicted state:

$$g_1\left(x\right) = x_1^2 + x_2^2; \tag{28}$$

$$g_2\left(x\right)\& = \left(x_1 - 1\right)^2 + \left(x_2 - 1\right)^2. \tag{29}$$

Minimizing g_1 attracts the state toward the origin, whereas minimizing g_2 attracts it toward (1, 1). The Pareto-optimal set in decision/state space is analytically known as

$$\mathcal{P}_x = \left\{\left(t,t\right) \middle| t \in \left[0,1\right]\right\}, \tag{30}$$

which provides an exact reference for measuring convergence in both state and objective spaces.

At each time step, an approximate Pareto set over horizon-H admissible control sequences $H = (u_0, \ldots, u_{H-1})$ is constructed using NSGA-II under constraints. Each candidate is evaluated by:

$$\begin{gathered} J\left(U; x_k\right) = \begin{bmatrix} g_1\left(\Phi_H\left(x_k, U\right)\right) \\ g_2\left(\Phi_H\left(x_k, U\right)\right) \end{bmatrix}; \\ S\left(U; u_{k-1}\right) = \sum_{i=0}^{H-1} \left\| u_i - u_{i-1} \right\|_2^2. \end{gathered} \tag{31}$$

From this candidate set, four selection strategies are compared:

1. WS: fixed weighted sum scalarization.
2. CHEB: Chebyshev scalarization relative to a utopia point.
3. PROJ: oracle baseline selecting the candidate whose terminal state is closest to $\mathcal{P}_x$.
4. LPF (ours): learned Pareto filter using a graph neural network.

Only the first control action of the selected candidate is applied in receding-horizon fashion.

Convergence is measured in both decision and objective spaces. The decision-space distance to the Pareto set is

$$d_x(x) = \min_{p \in \mathcal{P}_x} \|x - p\|_2, \tag{32}$$

while objective-space proximity is measured by distance to a dense sampling of the true front $\mathcal{P}_f$:

$$d_f(f) = \min_{q \in \mathcal{P}_f} \|f - q\|_2, \quad f = g(x). \tag{33}$$

Closed-loop simulations are run for T steps from the same initial condition x_0, using identical NSGA-II parameters across all strategies. Training data are generated from simulated rollouts. At each step, NSGA-II provides a candidate Pareto set. For each candidate, a teacher target is computed based on one-step progress toward $\mathcal{P}_x$:

$$y_k^{(i)} = d_x\left(x_k + T_s u_0^{(i)}\right) + \alpha \left\| u_0^{(i)} - u_{k-1} \right\|_2^2. \tag{34}$$

A graph representation of each candidate, encoding objective values, smoothness, and contextual state information, is used to train a GNN to regress this target via mean-squared error. During data generation, the oracle PROJ selector is used to ensure coverage of relevant regions of the state space. Table 1 summarizes the convergence statistics obtained from the closed-loop simulations.

Pareto convergence statistics over $T = 40$ steps.

Table 1

Pareto convergence statistics over $T = 40$ steps

Method	Final d_x	Final d_f	Mean d_x	Mean d_f
LPF	0.314	0.642	**0.405**	**0.663**
WS	**0.242**	**0.072**	0.751	1.280
CHEB	1.119	1.455	1.082	1.997
PROJ	0.278	0.767	0.609	1.221

Objective-space trajectories during the first control steps for all selectors, overlaid on the true Pareto front. LPF rapidly aligns its trajectory with the Pareto front, closely matching the oracle-like PROJ behavior, while WS and CHEB deviate due to fixed scalarization biases.

Fig. 1 provides a geometric interpretation of these results. While all strategies start from the same initial condition, their trajectories in objective space differ markedly. LPF rapidly steers the system toward the Pareto front and then evolves along it.

MPC-based control of boiler-turbine unit

To stress-test our strategy, we design the following experiment as a strict ablation of the decision stage in Pareto-front based nonlinear MPC for the boiler-turbine unit. All components of the control framework are kept identical to those reported in [27] except for the rule used to select a single control sequence from the Pareto front. We also decrease the Genetic Algorithm (GA) parameters to keep the Pareto front estimation in its early stages. This helps reduce calculation time but degrades the efficiency of the control strategy, ensuring that any observed performance difference is solely attributable to the Pareto front selection mechanism.

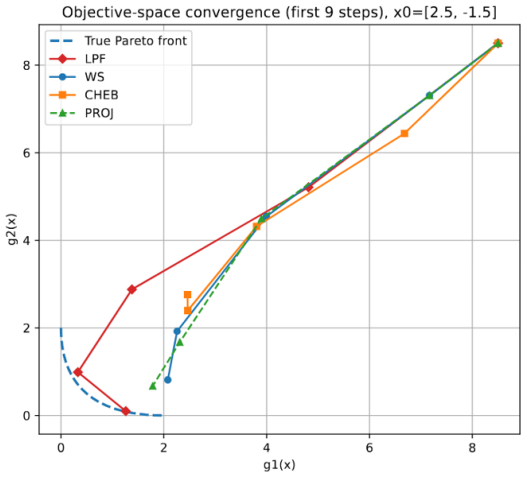


Fig. 1. Objective-space trajectories during the first control steps for all selectors, overlaid on the true Pareto front

The nonlinear model of a boiler-turbine unit is used without approximation or simplification. The state dynamics and the algebraic output relations for $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\boldsymbol{y}_3$ are preserved exactly as reported in [27]. The sampling time is $\boldsymbol{T}_s = 0.1$ s. The constraints of magnitude $0 \le u_q \le 1$, $q = 1, 2, 3$ and rate:

$$-0.007 \le \dot{u}_1 \le 0.007, \quad -2 \le \dot{u}_2 \le 2, \quad -0.05 \le \dot{u}_3 \le 0.05,$$

are enforced online through a trimming operator applied to every control update. The system is initialized at the 120 MW operating point:

$$x_{120} = \left[137,\ 120,\ 299.09\right]^T, \quad u_{120} = \left[0{,}563,\ 0.868,\ 0.743\right]^T,$$

and the reference is switched at one quarter of the simulation time to the 140 MW operating point:

$$x_{140} = \left[148.94,\ 140.7,\ 356.3\right]^T.$$

The total simulation time is 400 s. At each policy update, a multi-objective GA searches over a prediction horizon of $H = 10$ steps. The objectives are defined as the discounted absolute tracking errors of the three outputs along the horizon,

$$J_i = \sum_{k=0}^{H-1} \gamma^k \left| y_i(k) - y_i^{des} \right|, \quad i = 1,2,3,$$

with $\gamma = 0.9$. The GA parameters match those of [27]: population size 80, 40 generations, and NSGA-II non-dominated sorting.

Figs. 2 and 3 show the system outputs and control inputs during the load transition. We see that, under the same configurations, the baseline GA fails to control the system near its singular point, and causes a huge mathematical derived from the system equations. As this drift might not be physically accurate, the algorithms are working with the mathematical model that indeed has singular points in the dynamic model. Overcoming this drift is possible by increasing the horizon of the GA, as well as its genetic capability, like population and number of generations. However, our proposed LPF was able to avoid the singularity trap under the same configurations, showing a more stable performance. The main conclusions are:

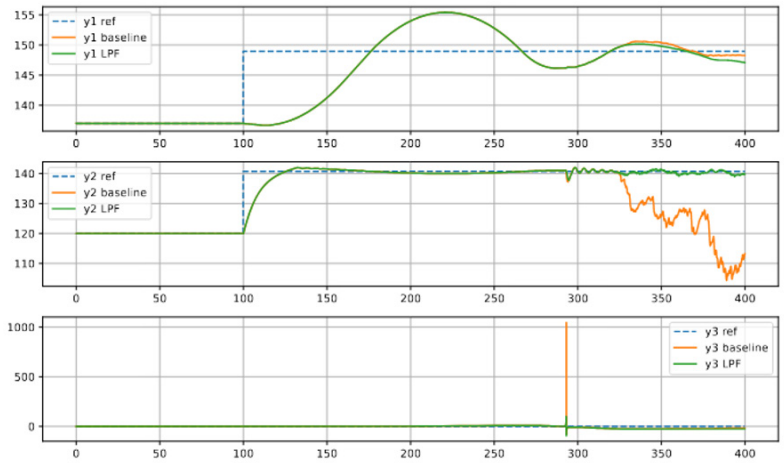


Fig. 2. Comparison between baseline GA selection and LPF-based selection during the MW to MW load transition. The LPF controller achieves smoother inputs and improved tracking performance

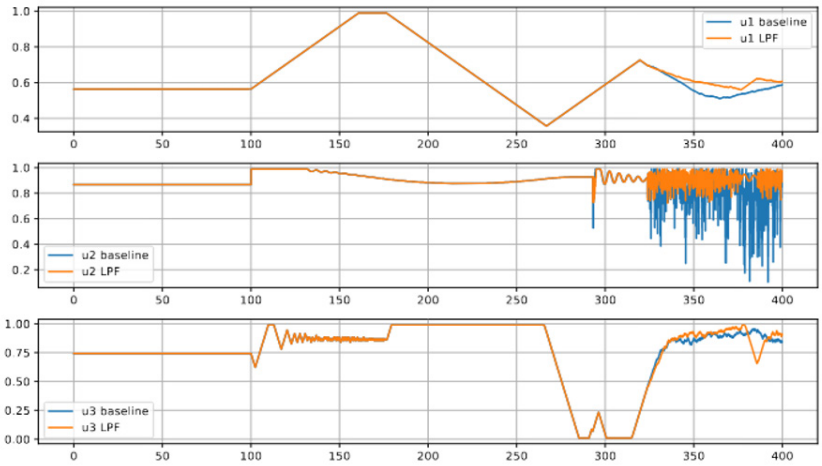


Fig. 3. Comparison between the control signals of the two algorithms

• The LPF-based controller produces significantly smoother control actions, avoiding the oscillatory behavior often selected by the baseline GA policy.

• The tracking of drum pressure, water level, and power output is faster and exhibits reduced overshoot with LPF.

• The LPF policy maintains the system inside the physically valid operating region of the boiler model, preventing the undesirable excursions that appear with purely objective-based selection.

These results demonstrate that the LPF does not alter the optimization process itself, but rather improves the decision stage by selecting physically consistent trajectories from the Pareto front. This confirms that structural awareness of candidate solutions leads to better closed-loop performance than traditional objective-based ranking, even when both methods use the same optimization backbone.

Conclusion

This paper introduced a structure-aware DSS for MOO-driven control under hard magnitude and slew-rate constraints. The core contribution lies in explicitly addressing the meta-decision gap inherent in Pareto-based receding-horizon control: while multi-objective optimizers provide sets of nondominated candidates, practical implementation requires selecting a single admissible control action at each time step. By decoupling Pareto-front generation from action selection, the proposed framework preserves the expressive power of MOO while replacing fragile fixed scalarization rules with a learned, structure-aware selector. The LPF leverages both candidate-specific attributes and relational information within the Pareto set, enabling context-dependent trade-off selection without manual weight tuning. Importantly, the LPF operates as a lightweight module, making it suitable for real-time constrained control architectures.

Experimental results on an analytically tractable benchmark demonstrated that the learned selector consistently guides the system toward the true Pareto set more reliably than classical weighted-sum or norm-based scalarizations. In the nonlinear boiler-turbine load-change scenario, the LPF-based controller achieved smoother control inputs, improved tracking performance, and avoided singular or unstable regions encountered by purely objective-based selection, despite identical optimization backbones. These findings confirm that, in Pareto-based control, the primary performance bottleneck often lies not in Pareto-front approximation but in the selection stage.

Future work will focus on extending the selector to provide theoretical closed-loop guarantees, integrating stability-aware training objectives, and exploring transferability across operating regimes and system classes. Additionally, incorporating uncertainty-awareness and adaptive online refinement of the selector may further enhance robustness in large-scale industrial applications.

REFERENCES

1. IEA. *World Energy Outlook 2022*, Available: Paris https://www.iea.org/reports/world-energy-outlook-2022 (Accessed 06.07.2026).

2. **Röth C., Milde F., Trebbels D., Schmidt J., Doppelbauer M.** A Stator With Offset Segments and a Double Stator Design for the Reduction of Torque Ripple of a Switched Reluctance Motor. *IEEE Transactions on Energy Conversion*. 2022, Vol. 37, No. 2, Pp. 1233–1240.

3. **Wang C., Liu M., Zhao Y., Qiao Y., Chong D., Yan J.** Dynamic modeling and operation optimization for the cold end system of thermal power plants during transient processes. *Energy*, 2018, Vol. 145, Pp. 734–746. DOI: 10.1016/j.energy.2017.12.146

4. **Ogata K.** *Modern control engineering*, 5th ed., New Jersey: Pearson Education, Inc., 2020.

5. **Schwenzer M., Ay M., Bergs T., Abel D.** Review on model predictive control: An engineering perspective. *The International Journal of Advanced Manufacturing Technology*, 2021, Vol. 117, No. 5, Pp. 1327–1349. DOI: 10.1007/s00170-021-07682-3

6. **Wu X., Shen J., Li Y., Lee K.Y.** Stable model predictive control based on TS fuzzy model with application to boiler-turbine coordinated system. *2011 50th IEEE Conference on Decision and Control and European Control Conference*, 2011, Pp. 2981–2987. DOI: 10.1109/CDC.2011.6160553

7. **Wu X., Shen J., Li Y., Lee K.Y.** Hierarchical optimization of boiler-turbine unit using fuzzy stable model predictive control. *Control Engineering Practice*, 2014, Vol. 30, Pp. 112–123. DOI: 10.1016/j.conengprac.2014.03.004

8. **Ławryńczuk M.** Nonlinear predictive control of a boiler-turbine unit: A state-space approach with successive on-line model linearisation and quadratic optimization. *ISA Transactions*, 2017, Vol. 67, Pp. 476–495. DOI: 10.1016/j.isatra.2017.01.016

9. **Yang C., Zhang T., Zhang Z., Sun L.** MLD–MPC for ultra-supercritical circulating fluidized bed boiler unit using subspace identification. *Energies*, 2022, Vol. 15, No. 15, Art. no. 5476. DOI: 10.3390/en15155476

10. **Banzhaf W., Machado P., Zhang M.** *Handbook of evolutionary machine learning*, NY: Springer, 2023.

11. **Lawrence N.P., Damarla S.K., Kim J.W., Tulsyan A., Amjad F., Wang K., Chachuat B., Lee J.M., Huang B., Bhushan Gopaluni R.** Machine learning for industrial sensing and control: A survey and practical perspective. *Control Engineering Practice*, 2024, Vol. 145, Art. no. 105841. DOI: 10.1016/j.conengprac.2024.10-5841

12. **Chuadhry M.A., M.R. G.R., Mathur A.P.** Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems. *Proceedings of the 6th ACM on Cyber-Physical System Security Workshop*, 2020, Pp. 23–29. DOI: 10.1145/3384941.340958

13. **Slowik A., Kwasnicka H.** Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications*, 2020, Vol. 32, Pp. 12363–12379. DOI: 10.1007/s00521-020-04832-8

14. **Deeb A., Khokhlovskiy V.N., Shkodyrev V.P.** Model predictive control and genetic algorithms for optimization of continuous stirred tank reactors. In: *Smart electromechanical systems* (eds. I.L. Tarasova, B.A. Kulik), 2024, Vol. 544, Pp. 185–191. DOI: 10.1007/978-3-031-64277-7_14

15. **Biegler L.T.** A perspective on nonlinear model predictive control. *Korean Journal of Chemical Engineering*, 2021, Vol. 38, No. 7, Pp. 1317–1332. DOI: 10.1007/s11814-021-0791-7

16. **Jones D.F., Florentino H.O.** Multi-objective optimization: Methods and applications. In: *The Palgrave handbook of operations research* (eds. S. Salhi, J. Boylan), 2022, Pp. 181–207. DOI: 10.1007/978-3-030-96935-6_6

17. **Deb K., Pratap A., Agarwal S., Meyarivan T.** A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 2002, Vol. 6, No. 2, Pp. 182–197. DOI: 10.1109/4235.996017

18. **Wang L., Cai Y., Ding B.** Robust model predictive control with bi-level optimization for boiler-turbine system. *IEEE Access*, 2021, Vol. 9, Pp. 48244–48253. DOI: 10.1109/ACCESS.2021.3066371

19. **Köhler J., Soloperto R., Müller M.A., Allgöwer F.** A computationally efficient robust model predictive control framework for uncertain nonlinear systems. *IEEE Transactions on Automatic Control*, 2020, Vol. 66, No. 2, Pp. 794–801. DOI: 10.1109/TAC.2020.2982585

20. **Wei Q., Liu Y., Lu J., Ling J., Luan Z., Chen M.** A new integral critic learning for optimal tracking control with applications to boiler-turbine systems. *Optimal Control: Applications and Methods*, 2023, Vol. 44, No. 2, Pp. 830–845. DOI: 10.1002/oca.2792

21. **Peitz S., Dellnitz M.** A survey of recent trends in multiobjective optimal control – surrogate models, feedback control and objective reduction. *Mathematical and Computational Applications*, 2018, Vol. 23, No. 2, Art. no. 30. DOI: 10.3390/mca23020030

22. **Bemporad A., de la Peña D.M.** Multiobjective model predictive control. *Automatica*, 2009, Vol. 45, Pp. 2823–2830. DOI: 10.1016/j.automatica.2009.09.032

23. **Zavala V.M., Flores-Tlacuahuac A.** Stability of multiobjective predictive control: A utopia-tracking approach. *Automatica*, 2012, Vol. 48, No. 10, Pp. 2627–2632. DOI: 10.1016/j.automatica.2012.06.066

24. **He D., Wang L., Sun J.** On stability of multiobjective NMPC with objective prioritization. *Automatica*, 2015, Vol. 57, Pp. 189–198. DOI: 10.1016/j.automatica.2015.04.024

25. **Grüne L., Stieler M.** Multiobjective model predictive control for stabilizing cost criteria. *Discrete and Continuous Dynamical Systems – B*, 2019, Vol. 24, No. 8, Pp. 3905–3928. DOI: 10.3934/dcdsb.2018336

26. **Xu J., Tian Y., Ma P., Rus D., Sueda S., Matusik W.** Prediction-guided multi- objective reinforcement learning for continuous robot control. *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 10607–10616.

27. **Deeb A., Khokhlovskiy V., Shkodyrev V.** Hierarchical multi-objective control of nonlinear systems with dynamical input constraints. *Artificial Intelligence and Applications*, 2026, Vol 4., No. 1., Pp. 57–67. DOI: 10.47852/bonviewAIA52024314

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Ali Deeb
Диб Али
E-mail: dib_a@spbstu.ru

Vladimir N. Khokhlovskiy
Хохловский Владимир Николаевич
E-mail: 78v.kh77@gmail.com

Submitted: 28.02.2026; Approved: 14.05.2026; Accepted: 16.06.2026.
Поступила: 28.02.2026; Одобрена: 14.05.2026; Принята: 16.06.2026.

Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 109–122.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 109–122.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19210
UDC 519.8

ADAPTIVE WEIGHT ADJUSTMENT IN MULTICRITERIA CONTROL OF NONLINEAR SYSTEMS BASED ON A LYAPUNOV FUNCTION

M.I. Fershtadt ✉ , *V.N. Shashikhin*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ tral1930@mail.ru

Abstract. The paper develops an approach to the control of nonlinear dynamic systems within a multicriteria optimization framework in which criterion weights are adjusted during the control process. The method relies on a Lyapunov function that is included both in the performance functional and in trajectory constraints. A numerical algorithm is proposed for dynamic tuning of the criterion weights. The criteria are stability and control energy consumption, while their relative importance changes depending on the current system state and trajectory. Adaptation is performed based on the current value of the Lyapunov function and the system state, which enables automatic prioritization of objectives without operator intervention. A numerical analysis of the system behavior under different weight-adaptation schemes is carried out. For the considered models, the simulation results show that adaptive weight adjustment can provide faster decay of the Lyapunov function and lower control energy consumption compared with fixed scalarization. The method can be applied to control problems for nonlinear and multi-agent systems, including cyber-physical complexes, autonomous agents, and energy nodes.

Keywords: multicriteria optimization, Lyapunov function, adaptive weights, nonlinear systems, stability

Citation: Fershtadt M.I., Shashikhin V.N. Adaptive weight adjustment in multicriteria control of nonlinear systems based on a Lyapunov function. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 109–122. DOI: 10.18721/JCSTCS.19210

© Fershtadt M.I., Shashikhin V.N., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19210
УДК 519.8

АДАПТИВНАЯ НАСТРОЙКА ВЕСОВ В ЗАДАЧЕ МНОГОКРИТЕРИАЛЬНОГО УПРАВЛЕНИЯ НЕЛИНЕЙНЫМИ СИСТЕМАМИ С ИСПОЛЬЗОВАНИЕМ ФУНКЦИИ ЛЯПУНОВА

М.И. Ферштадт ✉ , *В.Н. Шашихин*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ tral1930@mail.ru

Аннотация. В статье рассматривается подход к управлению нелинейными динамическими системами с использованием метода многокритериальной оптимизации, в котором веса критериев адаптируются в процессе управления. В основе метода лежит функция Ляпунова, которая входит как в функционал качества, так и выступает в качестве ограничения на траекторию системы. Предлагается численный алгоритм, позволяющий динамически настраивать веса критериев. Критериями являются устойчивость и энергозатраты, а их относительная значимость изменяется в зависимости от текущего состояния системы и траектории. Адаптация осуществляется на основе текущего значения функции Ляпунова и состояния системы, что обеспечивает автоматическую приоритизацию целей без вмешательства оператора. Проведен численный анализ поведения системы при различных схемах адаптации весов. Для рассмотренных моделей результаты численного моделирования показывают, что адаптивная настройка весов может обеспечивать более быстрое затухание функции Ляпунова и снижение энергозатрат по сравнению с фиксированной скаляризацией. Метод применим в задачах управления мультиагентными и нелинейными системами, включая киберфизические комплексы, автономные агенты и энергетические узлы.

Ключевые слова: многокритериальное управление, функция Ляпунова, адаптивные веса, нелинейные системы, устойчивость

Для цитирования: Fershtadt M.I., Shashikhin V.N. Adaptive weight adjustment in multicriteria control of nonlinear systems based on a Lyapunov function // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 109–122. DOI: 10.18721/JCSTCS.19210

Introduction

In modern control problems of complex systems, it is often necessary to account for several criteria simultaneously: stability, efficiency, coordination among subsystems, and energy consumption. Such problems belong to the class of multi-criteria optimization problems, whose solution requires balancing mutually competing objectives. This is especially relevant for nonlinear and multi-agent systems, where the dynamics may be unstable, while the control inputs may be constrained both by admissible magnitudes and by available information.

In practice, one of the most common approaches is to scalarize the vector objective function, for example, as a weighted sum of criteria. However, this leads to a central question: whether the criterion weights should be prescribed in advance or updated during the control process.

The theoretical foundations of optimization and decision-making in the problem of multi-criteria control have been described in classical works on mathematical programming, numerical optimization, system analysis, and control theory [1–4, 21, 24–26]. In particular, the development of linear and nonlinear programming, numerical optimization algorithms, and multi-criteria optimization methods

© Ферштадт М.И., Шашихин В.Н., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

is associated with the works of L. V. Kantorovich, G. B. Dantzig, L. G. Khachiyan, R. Steuer, D. G. Luenberger, S. Boyd, L. Vandenberghe, V. G. Zhadan, I. G. Chernorutskiy, N. Yu. Prokopenko, V. N. Kozlov, N. V. Rostov, and other authors [1–4]. In these works, multi-criteria problems are usually considered from the point of view of Pareto optimality, scalarization, criterion convolution, minimax schemes, penalty functions, or numerical methods for finding compromise solutions [1–4, 21, 24–26].

Weighted sum scalarization is one of the most common methods for converting a vector criterion into a scalar objective function. In this approach, individual criteria are multiplied by weight coefficients, and then standard numerical optimization algorithms can be applied to the resulting scalar problem [1–4]. However, the result of such scalarization depends significantly on the chosen weighting factors [21, 26]. If the weighting factors are fixed before the control process begins, they only reflect the developer's a priori preferences and do not account for changes in the relative importance of stability and control energy throughout the system's trajectory.

In the theory of control, multi-criteria optimization of parameters is used to configure regulators of nonlinear automatic control systems [2, 5, 6, 9, 10, 27]. For example, I. G. Chernorutskiy discusses optimization methods in the theory of control, including parametric optimization, multi-criteria problems, and Pareto compromise [2]. N. V. Rostov proposes methods for multi-criteria optimization of digital regulator parameters for stabilization and tracking systems with electromechanical installations, taking into account nonlinearities, constraints, and external disturbances [5, 6, 27–30]. In these studies, the main goal is to determine technically sound compromise or Pareto optimal regulator parameters. In this case, the weight coefficients of the individual criteria are usually selected as adjustable parameters and are not recalculated during the control process depending on the current state of the nonlinear system [5, 6, 25, 26].

Similar results have also been obtained in the field of distributed predictive control and optimal consensus for nonlinear multi-agent systems [7, 8, 14, 15, 20, 22, 23]. A. N. Venkat has developed a distributed predictive control system aimed at improving the performance and stability of closed-loop systems for large-scale interacting systems [7]. Y. Tang and X. Wang have considered optimal output consensus for heterogeneous nonlinear multi-agent systems with uncertainty, combining distributed optimization, consensus, and adaptive control mechanisms [8]. These approaches aim to ensure stability, constraint satisfaction, coordination, and distributed optimization, but they are based on predictive optimization, consensus protocols, and adaptive uncertainty compensation [7, 8, 14, 15, 20, 22, 23]. These approaches do not use the Lyapunov function as a state-dependent variable that directly redistributes the weights of the stability and energy criteria during the transient state [14–18, 20].

Therefore, the approach proposed in this article differs from existing results in the following ways. The goal is not to construct the entire Pareto set, and it is not just about a posteriori stability checking. The Lyapunov function is used both as a stability criterion and as a feedback variable that adaptively changes the weights of the criteria during control [9, 10, 14, 16–18, 20]. This allows for higher priority of stabilization when the system is far from equilibrium, and increased relative importance of energy saving near the equilibrium state.

Fixed weights do not allow the system to adapt to changing conditions. At the initial stage of the control process, it may be critical to ensure rapid decay of the system state (stability), while at later stages, reducing control costs becomes more important. Without proper adaptation, such changes in priority are not reflected in the control law, leading to suboptimal behavior.

This article proposes an approach in which the weights in the multi-objective control problem are determined adaptively, as functions of the current system state and the observed trajectory behavior. A distinctive feature of the approach is the use of the Lyapunov function both as a stability-related criterion and as a variable that governs the redistribution of priorities. This enables the stability criterion to receive greater weight automatically when approaching the boundaries of the phase space and its weakening when the system is already close to equilibrium, while maintaining control over costs.

Thus, the article solves two problems: it proposes an adaptive scalarization mechanism for the objective function and evaluates its effectiveness for constrained nonlinear control and for a multi-agent configuration.

Problem statement for multicriteria control with adaptive weights

Consider a nonlinear controlled system in the form of ordinary differential equations:

$$\dot{x}(t) = f\left(x(t), u(t)\right);$$
$$x(0) = x_0;$$
$$x(t) \in X;$$
$$u(t) \in U,$$

where $x(t) \in \mathbb{R}^n$ – the state vector; $u(t) \in \mathbb{R}^n$ – control input; $f(\ ,\ )$ – a smooth vector function that defines the dynamics of the system; X, U – the admissible sets of states and control inputs, respectively.

For application to real objects, the proposed method assumes that the state vector x is available for feedback control. This means that all components of x are either directly measured by sensors or reconstructed with sufficient accuracy using a state observer. This article focuses on the case of full state feedback. The synthesis of observers for partially measured systems is beyond the scope of this work.

The proposed approach is applicable to nonlinear controlled objects for which: a state-space model is known in the admissible operating domain; a continuously differentiable positive-definite Lyapunov function $V(x)$ can be constructed; the admissible set of controls is non-empty; and the available control actions allow for the reduction of the function $V(x)$. Such objects include electromechanical systems, robotic modules, active stabilization systems, power electronic converters, and homogeneous groups of mobile agents, provided that the necessary state variables are measurable [10, 14, 18–20, 22, 27].

In the following, by the Lyapunov function $V(x)$, we mean a continuously differentiable positive-definite function such that

$$V(0) = 0, \quad V(x) > 0, \quad x \neq 0.$$

The value of $V(x)$ is used as a quantitative measure of the system's departure from equilibrium, and a decrease in $V(x)$ along the trajectory indicates the system's stabilizing behavior. In simulations, a quadratic function is used for a two-dimensional model:

$$V(x) = x_1^2 + x_2^2.$$

For other nonlinear systems, the form of the Lyapunov function may differ and should be determined based on the structure of the specific model. Thus, the proposed approach is intended for nonlinear controlled systems for which the Lyapunov function is known or can be constructed in the admissible state domain, and for which there exists an admissible control input that can reduce this function, at least in the sense of the penalty used in the numerical algorithm. In the case of multiple agents, discussed below, it is assumed that the agents are homogeneous, the interaction graph is connected, and the agents share a common goal of achieving equilibrium. Systems with heterogeneous agent dynamics or incompatible local goals require additional assumptions and are not considered in this work.

The goal of control is to minimize a vector criterion that includes at least two individual criteria:

– the criterion of stability $J_1(u) = \int_0^T V\left(x(t)\right) dt$;

– control cost criteria $J_2(u) = \int_0^T \|u(t)\|^2 dt$.

Scalarization form with fixed weights:

$$J(u) = \int_0^T \left(\alpha V(x(t)) + \beta \|u(t)\|^2\right) dt,$$

where $\alpha > 0$ and $\beta > 0$; α and β are predefined parameters that specify the relative priority of the objectives.

Fixed values of α and β do not adapt to the system dynamics and do not reflect the current importance of the criteria at different points along the trajectory.

Adaptive weight structure

It is proposed to define weights as functions of the current state:

$$\alpha(t) = \phi_1(x(t));$$
$$\beta(t) = \phi_2(x(t)),$$

where ϕ_1 and ϕ_2 are smooth functions with the following properties:

– ϕ_1 increases with $V(x)$: the stability-related component receives greater weight as the system moves farther from equilibrium;

– ϕ_2 increases as we approach equilibrium: the control-cost component becomes more influential as the state approaches equilibrium.

The specified properties do not set the only possible form of weight functions, but define a class of admissible adaptive scalarizations [21, 25, 26]. In general, any smooth positive functions $\phi_1(V)$, $\phi_2(V)$ can be used, for which the normalized weight of the stability criterion:

$$\omega_1(V) = \frac{\varphi_2(V)}{\varphi_1(V) + \varphi_2(V)},$$

does not decrease as $V(x)$ increases, and the normalized weight of the energy consumption criterion:

$$\omega_2(V) = \frac{\varphi_2(V)}{\varphi_1(V) + \varphi_2(V)},$$

does not increase with $V(x)$. Therefore, the proposed below form of weights is considered not as the only possible, but as a simple numerically implementable option, allowing to investigate the effect of dynamic redistribution of priorities between stability and energy saving.

Taking into account the adaptive weights, a generalized scalarized functional is introduced:

$$J(u) = \int_0^T \left[\phi_1(x(t)) \cdot V(x(t)) + \phi_2(x(t)) \cdot \|u(t)\|^2\right] dt.$$

An example of admissible weight functions is

$$\phi_1(x) = 1 + \kappa_1 \cdot V(x);$$
$$\phi_2(x) + 1 + \kappa_2 \cdot \frac{1}{1 + V(x)},$$

where $\kappa_1, \kappa_2 \geq 0$ are the adaptation coefficients.

This construction automatically increases the priority of stabilization in the early stages of control (when $V(x)$ is large) and increases the relative importance of reducing control energy when $V(x) \to 0$.

The specific form of the functions ϕ_1, ϕ_2 and the values of the parameters κ_1, κ_2 affect the quantitative characteristics of the transient process. Increasing κ_1 enhances the stabilizing priority for large deviations, which can reduce the convergence time but also increase the initial control input and lead to saturation of the u constraint. Increasing κ_2 increases the importance of energy saving near the equilibrium, which reduces the integral energy consumption but may slow down the final convergence phase. Therefore, the numerical results should be interpreted as a result for the chosen parametric family of weight functions, rather than as a universal statement for all possible adaptive dependencies [21, 26].

Limitations of stability

In addition to including $V(x)$ in the objective function, the following condition may be imposed as a constraint:

$$\dot{V}\left(x(t)\right) = \nabla V\left(x(t)\right)^T f\left(x(t), u(t)\right) \le -\delta \left\| x(t) \right\|^2, \quad \delta > 0.$$

If this condition is imposed as a hard constraint, monotone decrease of the Lyapunov function is enforced. If the condition is introduced through a penalty term, this decrease is promoted but is not guaranteed by construction.

Discretization of the Problem

For numerical implementation, the problem is discretized with respect to time:

$$x_{k+1} = f_d\left(x_k, u_k\right), \quad k = 0, \ldots, N-1,$$

where f_d is a discrete model of the system obtained, for example, by the Euler method:

$$x_{k+1} = x_k + \Delta t \cdot \left(x_k, u_k\right),$$

or by a Runge–Kutta method.

The scalarized functional is then written in discrete form:

$$J(u) = \sum_{k=0}^{N-1} \left[\phi_1\left(x_k\right) \cdot V\left(x_k\right) + \phi_2\left(x_k\right) \cdot \left\| u_k \right\|^2 \right] \cdot \Delta t.$$

Numerical Implementation of the Algorithm with Adaptive Weights

Consider the numerical implementation of the proposed approach to multi-objective control with adaptive tuning of the functional weights. The problem is formulated as a multi-objective problem (stability and energy consumption), and a feedback law is then introduced, in which the adaptive gain $G(x_k)$ is determined based on the weights $\phi_1(x_k)$, $\phi_2(x_k)$. The functional J is used as a quality criterion and for selecting the controller structure and weight adaptation mechanism. In this case, the minimization of J does not have to be repeated at every control step.

Solution algorithm

1. **Initialization of parameters**:

Initial state $x_0 \in \mathbb{R}^n$;

Simulation horizon T and sampling step $\Delta t = \frac{T}{N}$;

Admissible bounds on the control input $u_k \in \left[-u_{\max}, u_{\max}\right]$;

The Lyapunov function $V\left(x\right) = \left\|x\right\|^2$;

Adaptive weights $\phi_1(x_k)$, $\phi_2(x_k)$.

2. **Construction of the objective function**:

$$J\left(u\right) = \sum_{k=0}^{N-1}\left[\left(1+\kappa_1 V\left(x_k\right)\right)\cdot V\left(x_k\right) + \left(1+\kappa_2\cdot\frac{1}{1+V\left(x_k\right)}\right)\cdot\left\|u_k\right\|^2\right]\cdot\Delta t.$$

3. **The system dynamics may be integrated numerically by several scheme**s:

The simplest model: Euler method

$$x_{k+1} = x_k + \Delta t\cdot f\left(x_k, u_k\right).$$

For improved accuracy of trajectory and stability estimates, a fourth-order Runge–Kutta method may be used:

$$x_{k+1} = x_k + \frac{\Delta t}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right),$$

where $k_1 = f\left(x_k, u_k\right)$, $k_2 = f\left(x_k + \frac{\Delta t}{2}k_1, u_k\right)$, $k_3 = f\left(x_k + \frac{\Delta t}{2}k_2, u_k\right)$, $k_4 = f\left(x_k + \Delta t\cdot k_3, u_k\right)$.

4. **Constraint Handling**:

Control: $u_k \in \left[-u_{\max}, u_{\max}\right]$;

States: $x_k \in \left[-x_{\max}, x_{\max}\right]$;

The Lyapunov derivative:

$$\dot V\left(x_k\right) = \nabla V\left(x_k\right)^T f\left(x_k, u_k\right) = 2x_k^T f\left(x_k, u_k\right) \le -\delta\left\|x_k\right\|^2.$$

The constraints may be incorporated in two forms: either as soft penalties in the objective function by adding a term $\lambda\max\left\{0, \nabla V\left(x_k\right)^T f\left(x_k, u_k\right) + \delta\left\|x_k\right\|^2\right\}^2$, $\lambda > 0$, or as inequality constraints included directly in the optimization formulation.

Optimization

To implement the proposed approach, a feedback control scheme with an adaptively adjusted gain is used. At each step, the control u_k is calculated as follows:

1. The current adaptive weights $\phi_1(x_k)$ and $\phi_2(x_k)$ are computed from the system state;
2. The control input is defined as a function of the weight ratio:

$$u_{i,k} = -G\left(\phi_1\left(x_{i,k}\right), \phi_2\left(x_{i,k}\right)\right)\cdot\left(Kx_{i,k} + \rho\sum_{j\in N_i}\left(x_{i,k} - x_{j,k}\right)\right),$$

where G is an adaptive gain that balances the convergence rate (ϕ_1) and control-resource expenditure (ϕ_2).

Stabilization and regularization

To prevent divergence of the numerical procedure, the following methods are used:

1. Projection of variables onto the admissible domain;
2. Penalization of Lyapunov-function increase in the case of model inaccuracies (penalty for $\dot{V}(x_k) + \delta\|x_k\|^2 > 0$
3. Preventing oscillations at the equilibrium point (threshold-based setting of u_k to zero);
4. Filtering of numerical noise and low-amplitude oscillations (additional filtering).

Simulation of the Adaptive-Weight Method

To evaluate the effectiveness of the proposed approach with adaptive weight adjustment, numerical experiments were conducted for models with nonlinear dynamics to assess how adaptive weights affect system behavior, control energy, trajectory stability, and Lyapunov-function decay.

Experiment 1: A Single Nonlinear System

Model:

$$\begin{cases} \dot{x}_1 = -x_1 + x_2^2 + u \\ \dot{x}_2 = -x_2 + x_1 x_2 \end{cases}; \quad V(x) = x_1^2 + x_2^2.$$

Initial state: $x_0 = [1.5, -1.0]$;

Simulation horizon: $T = 15$, $N = 1500$, $\Delta t = 0.01$;

Constraints: $u_k \in [-1.5, 1.5]$, $x_k \in [-10, 10]$;

In experiment 1, the values $\kappa_1 = 5.0$, $\kappa_2 = 2.0$, $\delta = 0.05$ are used, which were selected based on the results of preliminary numerical experiments to ensure a pronounced priority of stabilization when the state is far from equilibrium and consistent redistribution of weights in the vicinity of the equilibrium state.

To evaluate the effectiveness, we compare the baseline method with fixed scalarization and the proposed method with adaptive weights. In both cases, we use the same initial conditions, control and state constraints, and the same numerical integration scheme.

In the fixed-weight baseline, the coefficients were set to constant values $\phi_1 = 2.0$, $\phi_2 = 3.0$. Therefore, the corresponding constant gain in the feedback law was

$$G = \frac{\varphi_1}{\varphi_1 + \varphi_2} = 0.4.$$

These values remained unchanged throughout the trajectory. In the adaptive case, the same feedback structure was used, but the coefficients $\phi_1(x_k)$ and $\phi_2(x_k)$ were recalculated at each time step as functions of the current value of the Lyapunov function. Both simulations used the same initial conditions, integration scheme, control bounds, state bounds, and Lyapunov-based correction procedure. Therefore, the comparison highlights the effect of replacing a constant priority ratio with state-dependent priority redistribution [1–4, 21, 26].

It should be noted that a specially tuned pair of constant weights may reduce the quantitative advantage of the adaptive approach [21, 26]. Therefore, the obtained results should not be interpreted as proof of superiority over all possible fixed-weight scalarizations.

Results

For the selected nonlinear model and the specified parameters, the Lyapunov function $V(x_k)$ demonstrates a higher convergence rate to zero when using adaptive weights compared with the selected

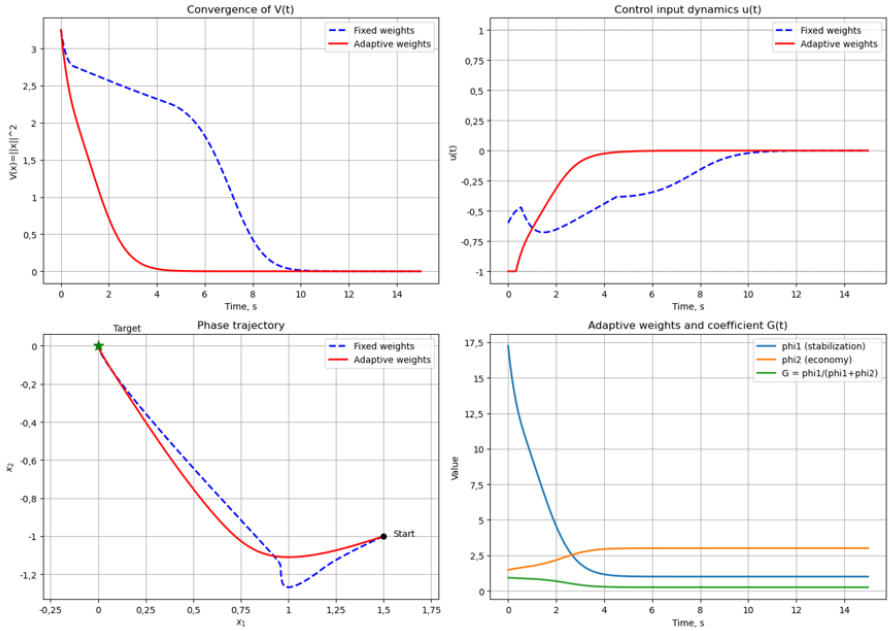


Fig. 1. Comparison of fixed-weight and adaptive-weight control for a single system: (a) – dependence $V(t)$; (b) – dependence $u(t)$; (c) – phase trajectory; (d) – change in weights ϕ_1, ϕ_2 and coefficient $G(t)$

fixed-weight scalarization baseline. This indicates a reduction in the transition time in this numerical experiment, but it does not necessarily generalize to all nonlinear systems;

In the simulation, the adaptive redistribution of priorities leads to a larger control input at the initial stage and to a decrease in it near the equilibrium. As a result, the integral control energy $E = \int_0^T u^2(t)\,dt$ is lower than that of the considered fixed-weight baseline. The magnitude of this reduction depends on the selected functions ϕ_1, ϕ_2, parameters κ_1, κ_2, control constraints, and the selected fixed-weight baseline level.

The phase trajectory also illustrates the effect of adaptive priority reallocation for the selected test model. This result should be interpreted as a numerical illustration of the proposed mechanism, rather than as a general proof of the superiority of arbitrary nonlinear systems. The adaptation mechanism is manifested in the fact that at the beginning of the process, the stabilizing component has a greater relative influence, while near equilibrium, the relative influence of the control cost component increases.

Experiment 2: A Multi-Agent Consensus System

Consider a distributed system of $M = 3$ agents. In this simulation, the agents are homogeneous: each agent has the same local dynamics, but different initial conditions. Let the state of the i-th agent be denoted by

$$x_i = \left(x_{1,i}, x_{2,i}\right)^T, \quad i = 1, \ldots, M.$$

The local dynamics of each agent are defined by the system

$$\dot{x}_{1,i} = -x_{1,i} + x_{2,i}^2 + u;$$
$$\dot{x}_{2,i} = -x_{2,i} + x_{1,i}x_{2,i}.$$

Thus, the control input ui enters the first equation of the agent's local dynamics. The connection between agents is introduced not as an additional physical term in the right-hand side of the system, but through a matching component in the control law.

The interaction graph is assumed to be fully connected. A normalized adjacency matrix is used for three agents [7, 8, 18, 20, 22]

$$A = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

The following adaptive control law is used:

$$u_{i,k} = -G\left(\varphi_{1,i}, \varphi_{2,i}\right)\left(Kx_{1,i,k} + \rho \sum_{j \in N_i} a_{ij}\left(x_{1,i,k} - x_{1,j,k}\right)\right),$$

where K determines the intensity of local agent stabilization, and ρ specifies the strength of the consensus interaction with neighboring agents; in the simulation under consideration, K = 1.0 and ρ = = 1.0 were used, a_{ij} are the elements of adjacency matrix, N_i is the set of neighbors of the i-th agent.

A local Lyapunov function is used for each agent:

$$V_i\left(x_i\right) = x_{1,i}^2 + x_{2,i}^2.$$

Weight functions:

$$\phi_1\left(x_i\right) = 1 + \kappa_1 \left\|x_i\right\|^2, \quad \phi_2\left(x_i\right) = 1 + \frac{\kappa_2}{1 + \left\|x_i\right\|^2}.$$

Thus, each agent adapts the local weights based on its own state, but the consensus coupling term in the control law takes into account the deviation from neighboring agents. The local decrease in $V_i(x_i)$ alone is not sufficient to prove the stability and coordination properties of the entire multi-agent system, so the consensus error and the average value of the Lyapunov function over the group are also analyzed [8, 18, 20, 22].

Constraints: $u_k \in [-1.5, 1.5], \quad x_k \in [-10, 10]$;

The parameters: κ_1 = 2.0, κ_2 = 10.0, δ = 0.05.

Heterogeneous agents with different local dynamics $f_i(x_i, u_i)$ are not considered in this article and require a separate analysis of the conditions for consensus and stability.

For the considered homogeneous system of three agents with a fully connected interaction graph, the simulation demonstrates a decrease in the disagreement between the agents and convergence of their states to a neighborhood of a common equilibrium.

In this experiment, the total control energy is reduced by more than 40% compared to the considered fixed-weight baseline. This effect is associated with an earlier reduction of the control effort near equilibrium. The quantitative value of the gain depends on the selected adaptive weight functions, the parameters κ_1, κ_2, the control constraints, and the selected fixed-weight baseline used for comparison.

The control inputs reach their highest values at the beginning of the transition process, and then decrease as the agent trajectories approach each other. In the simulation, the average Lyapunov function for the group tends to zero. This indicates a stabilization of the behavior of the chosen multi-

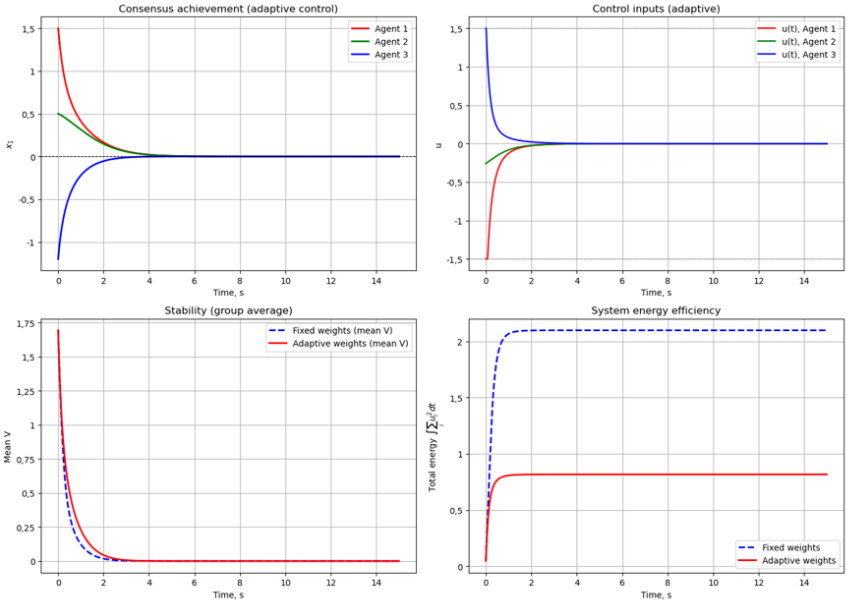


Fig. 2. Results of a numerical experiment for a multi-agent system:
(a) – consensus formation process; (b) – control inputs of the agents;
(c) – group dynamics of the Lyapunov function; (d) – total energy consumption

agent model, but this should be interpreted together with the consensus-error curve, since local Lyapunov reduction and network-level consensus are different properties of a distributed system.

The Python code to generate the plots (Figs. 1, 2) is available for download and can be run locally[1].

Conclusions

The conducted research shows that adaptive adjustment of the weight coefficients in a multi-criteria functional can be used as a control tool for nonlinear systems. The obtained results are applicable to models for which a Lyapunov function can be constructed, the state variables are measurable, control constraints are specified, and a stabilizing control law can be implemented. The main conclusions are as follows:

1. A mechanism is proposed in which the weight coefficients of the objective function are dynamically changed depending on the current state of the system through the Lyapunov function. Unlike classical weighted scalarization, in which the weight coefficients are fixed before the optimization or control process begins, the proposed approach allows the weight coefficients of the criteria to be recalculated in real-time. This allows for a shift in priority towards stabilization in the case of large deviations from equilibrium and towards energy conservation as the system approaches equilibrium. The chosen form of the weight coefficients is not the only possible choice and is considered as a representative of the class of smooth positive adaptive functions.

2. The proposed approach is a development of existing approaches to multi-criteria control by using the Lyapunov function not only as a stability criterion, but also as a variable that regulates the redistribution of priorities between individual criteria. This distinguishes this method from known approaches to multi-criteria parameter optimization, where the Pareto-optimal regulator parameters are typically selected with fixed-weight coefficients for the criteria [2, 5, 6, 21, 25, 26].

[1] GitHub – fershtadt/Adaptive-scale-adjustment · GitHub, Available: https://github.com/fershtadt/Adaptive-scale-adjustment (Accessed 21.07.2026)

3. A numerically implementable condition for reducing the Lyapunov function is introduced. This condition can be used either as an explicit constraint or as a penalty term in the quality functional. If there exists a valid control that satisfies this condition, it helps to bring the system to an equilibrium state. The fulfillment of this condition depends on the nonlinear model, the initial state, the set of valid controls, and the chosen Lyapunov function.

4. For the considered multi-agent system, where the communication between the subsystems is provided by a connected graph, it is shown by numerical methods that the local adaptation of weights based on the Lyapunov functions is compatible with the coordination term in the control law and leads to a reduction in the discrepancy between the agents. Unlike distributed model predictive control approaches, the proposed scheme does not require complex numerical optimization at each time step [7, 14, 15, 20, 23]. Instead, the redistribution of priorities is performed directly using the adaptive weights based on the Lyapunov functions.

5. The algorithm can be implemented in a centralized form and, under additional assumptions, in a distributed multi-agent form, provided that each agent has access to its own state and the states of neighboring agents in the graph. In this work, the multi-agent implementation was confirmed by numerical simulations for a system of three agents. The use of simple numerical integrators, such as the Euler method or the Runge-Kutta method, makes the scheme convenient for computational implementation.

6. The method was simulated in Python. Under the simulation parameters, the results demonstrate a faster decrease in the Lyapunov function compared to the selected baseline with fixed weights and a decrease in energy consumption near the system's equilibrium. The proposed method can be used to achieve a trade-off between fast stabilization and reduced control inputs [9, 10, 14, 16–18, 20]. These findings are applicable only to the class of systems under consideration. Future work will address the scaling of this approach to more complex real-world systems.

REFERENCES

1. **Zhadan V.G.** *Metody optimizatsii. Chast' II. Chislennye algoritmy* [*Optimization Methods. Part II. Numerical Algorithms*]. Moscow: MFTI, 2015.

2. **Chernorutskii I.G.** *Metody optimizatsii v teorii upravleniia* [*Optimization methods in control theory*]. St. Petersburg: Piter, 2004.

3. **Prokopenko N.Iu.** *Metody optimizatsii* [*Optimization methods*]. Nizhny Novgorod: NNGASU, 2018.

4. **Kochenderfer M.J., Wheeler T.A.** *Algorithms for Optimization*, 2nd ed. Cambridge, MA; London: MIT Press, 2025.

5. **Rostov N.V.** Sequential multiobjective parametrical optimization of controllers for nonlinear control system. St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems, 2010, Vol. 113, No. 6, Pp. 44–50.

6. **Rostov N.V.** Multiobjective parameter optimization of digital controllers with regard to the influence of nonlinearities and external disturbances. *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems*, 2014, Vol. 193, No. 2, Pp. 91–98.

7. **Venkat A.N.** *Distributed Model Predictive Control: Theory and Applications*, PhD dissertation. Madison: University of Wisconsin–Madison, 2006.

8. **Tang Y., Wang X.** Optimal Output Consensus for Nonlinear Multiagent Systems with Both Static and Dynamic Uncertainties. *IEEE Transactions on Automatic Control*, 2021, Vol. 66, No. 4. Pp. 1733–1740. DOI: 10.1109/TAC.2020.2996978

9. **Fershtadt M.I., Shashikhin V.N.** Multi-criteria control of large-scale nonlinear dynamical systems without linearization, based on Lyapunov functions. *Computing, Telecommunications and Control*, 2025, Vol. 18, No. 4, Pp. 112–122. DOI: 10.18721/JCSTCS.18410

10. **Shashikhin V.N., Budnik S.V., Golovina K.O.** Control of the spectrum of Lyapunov characteristic exponents in nonlinear large-scale systems. *Computing, Telecommunications and Control*, 2021, Vol. 14, No. 4, Pp. 37–51. DOI: 10.18721/JCSTCS.14404

11. **Zhang C., Fu J.** Multi-objective dynamic optimization of path-constrained switched systems. *Automatica*, 2024, Vol. 159, Art. no. 111326. DOI: 10.1016/j.automatica.2023.111326

12. **Tao M., Li Q., Yu J.** Multi-Objective Dynamic Path Planning with Multi-Agent Deep Reinforcement Learning. *Journal of Marine Science and Engineering*, 2024, Vol. 13, No. 1, Art. no. 20. DOI: 10.3390/jmse13010020

13. **Lin J., He C., Tian Y., Pan L.** Variable Reconstruction for Evolutionary Expensive Large-Scale Multiobjective Optimization and Its Application on Aerodynamic Design. *IEEE/CAA Journal of Automatica Sinica*, 2025, Vol. 12, No. 4, Pp. 719–733. DOI: 10.1109/JAS.2024.124947

14. **Zheng Y., Wang Y., Li S.** Adaptive control Lyapunov function based model predictive control for continuous nonlinear systems. *International Journal of Robust and Nonlinear Control*, 2022, Vol. 33, No. 2, Pp. 1254–1266. DOI: 10.1002/rnc.6409

15. **Zheng X.-Y., Yan H.-S.** Lyapunov-based stochastic model predictive control of stochastic nonlinear systems with input-delay using multi-dimensional Taylor network. *Journal of the Franklin Institute*, 2025, Vol. 362, No. 9, Art. no. 107694. DOI: 10.1016/j.jfranklin.2025.107694

16. **Baheri A.** Distributionally robust Lyapunov–Barrier Networks for safe and stable control under uncertainty. *Results in Control and Optimization*, 2025, Vol. 19, Art. no. 100556. DOI: 10.1016/j.rico.2025.100556

17. **Stiti C., Benrabah M., Aouaichia A., Oubelaid A., Bajaj M., Tuka M.B., Kara K.** Lyapunov-based neural network model predictive control using metaheuristic optimization approach. *Scientific Reports*, 2024, Vol. 14, Art. no. 18760. DOI: 10.1038/s41598-024-69365-9

18. **Nersesov S.G., Haddad W.M.** Control vector Lyapunov functions for large-scale impulsive dynamical systems. *Nonlinear Analysis: Hybrid Systems*, 2007, Vol. 1, No. 2, Pp. 223–243. DOI: 10.1016/j.nahs.2006.10.006

19. **Chen Y.-H.** Nonlinear Adaptive Optimal Control Design and Implementation for Trajectory Tracking of Four-Wheeled Mecanum Mobile Robots. *Mathematics*, 2024, Vol. 12, No. 24, Art. no. 4013. DOI: 10.3390/math12244013

20. **Yan Z., Zhang M., Zhou J., Yue L.** Distributed Lyapunov-Based Model Predictive Control for AUV Formation Systems with Multiple Constraints. *Journal of Marine Science and Engineering*, 2024, Vol. 12, No. 3, Art. no. 363. DOI: 10.3390/jmse12030363

21. **Peitz S., Dellnitz M.** A Survey of Recent Trends in Multiobjective Optimal Control—Surrogate Models, Feedback Control and Objective Reduction. *Mathematical and Computational Applications*, 2018, Vol. 23, No. 2, Art. no. 30. DOI: 10.3390/mca23020030

22. **Wen J., Li L., Wu Q., Li K., Lu J.** Multi-Objective Cooperative Adaptive Cruise Control Platooning of Intelligent Connected Commercial Vehicles in Event-Triggered Conditions. *Actuators*, 2024, Vol. 13, No. 12, Art. no. 522. DOI: 10.3390/act13120522

23. **Ionescu C.M., Caruntu C.F., Cajo R., Ghita M., Crevecoeur G., Copot C.** Multi-Objective Predictive Control Optimization with Varying Term Objectives: A Wind Farm Case Study. *Processes*, 2019, Vol. 7, No. 11, Art. no. 778. DOI: 10.3390/pr7110778

24. **El Sayed M.A., Farahat A.F., Elsisy M.A., Alsabaan M., Ibrahem M.I., Elwahsh H.** Two TOPSIS-Based Approaches for Multi-Choice Rough Bi-Level Multi-Objective Nonlinear Programming Problems. *Mathematics*, 2025, Vol. 13, No. 8, Art. no. 1242. DOI: 10.3390/math13081242

25. **Nelyubin A.P., Podinovski V.V.** Multicriteria Problems with Importance-Ordered Criteria Groups. *Automation and Remote Control*, 2022, Vol. 83, Pp. 1108–1122. DOI: 10.1134/S0005117922070074

26. **Podinovski V.V., Potapov M.A.** Analysis of the Sensitivity of Solutions of Multi-Criteria Problems Based on Parametric Partial Preference Relations. *Automation and Remote Control*, 2019, Vol. 80, Pp. 1294–1303. DOI: 10.1134/S0005117919070075

27. **Balandin D.V., Kogan M.M.** Multicriteria Robust Generalized H_2 and γ_0 Controllers with Application to Stabilization of a Rotor in Electromagnetic Bearings. *Automation and Remote Control*, 2018, Vol. 79, Pp. 996–1012. DOI: 10.1134/S0005117918060024

28. **Demidenko O.M., Borchik E.M., Yakimov A.I.** Multi-criterial optimization of resource distribution in the process of finished products. *Problems of Physics, Mathematics and Technics*, 2022, Vol. 52, No. 3, Pp. 90–96. DOI: 10.54341/20778708_2022_3_52_90

29. **Melnikov A.V., Karavaev A.A., Zheleznyakov A.O.** Analysis of the main approaches to modeling management processes for special-purpose organizational and technical systems. *Herald of Dagestan State Technical University. Technical Sciences*, 2024, Vol. 51, No. 1, Pp. 153–165. DOI: 10.21822/2073-6185-2024-51-1-153-165

30. **Belykh M.A.** Formalization of a multi-criteria transport task with time constraints. *Modeling, Optimization and Information Technology*, 2024, Vol. 12, No. 2, Pp. 1–9. DOI: 10.26102/2310-6018/2024.45.2.027

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Mikhail I. Fershtadt
Ферштадт Михаил Иосифович
E-mail: tral1930@mail.ru

Vladimir N. Shashikhin
Шашихин Владимир Николаевич
E-mail: shashihin@bk.ru

Submitted: 07.04.2026; Approved: 08.06.2026; Accepted: 25.06.2026.
Поступила: 07.04.2026; Одобрена: 08.06.2026; Принята: 25.06.2026.