DOI: 10.31618/ESU.2413-9335.2025.1.127

Евразийский Союз Ученых.
Серия: технические и физико-математические науки

Ежемесячный научный журнал
№ 2 (127)/2025 Том 1

ГЛАВНЫЙ РЕДАКТОР

Макаровский Денис Анатольевич
AuthorID: 559173
Заведующий кафедрой организационного управления Института прикладного анализа поведения и психолого-социальных технологий, практикующий психолог, специалист в сфере управления образованием.

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

•**Штерензон Вера Анатольевна**
AuthorID: 660374
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, Институт новых материалов и технологий (Екатеринбург), кандидат технических наук

•**Синьковский Антон Владимирович**
AuthorID: 806157
Московский государственный технологический университет "Станкин", кафедра информационной безопасности (Москва), кандидат технических наук

•**Штерензон Владимир Александрович**
AuthorID: 762704
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, Институт фундаментального образования, Кафедра теоретической механики (Екатеринбург), кандидат технических наук

•**Зыков Сергей Арленович**
AuthorID: 9574
Институт физики металлов им. М.Н. Михеева УрО РАН, Отдел теоретической и математической физики, Лаборатория теории нелинейных явлений (Екатеринбург), кандидат физ-мат. наук

•**Дронсейко Виталий Витальевич**
AuthorID: 1051220
Московский автомобильно-дорожный государственный технический университет (МАДИ), Кафедра "Организация и безопасность движения" (Москва), кандидат технических наук

Статьи, поступающие в редакцию, рецензируются. За достоверность сведений, изложенных в статьях, ответственность несут авторы. Мнение редакции может не совпадать с мнением авторов материалов. При перепечатке ссылка на журнал обязательна. Материалы публикуются в авторской редакции.

Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций.

Художник: Валегин Арсений Петрович
Верстка: Курпатова Ирина Александровна

Адрес редакции:
198320, Санкт-Петербург, Город Красное Село, ул. Геологическая, д. 44, к. 1, литера А
E-mail: info@euroasia-science.ru ;
www.euroasia-science.ru

Учредитель и издатель ООО «Логика+»
Тираж 1000 экз.

СОДЕРЖАНИЕ

МАТЕМАТИКА И МЕХАНИКА

КОМПЬЮТЕРНЫЕ НАУКИ И ИНФОРМАТИКА

ТЕХНИЧЕСКИЕ НАУКИ

МАТЕМАТИКА И МЕХАНИКА

АНАЛИЗ ВЛИЯНИЯ ОТДЕЛЬНЫХ АРЕНДАТОРОВ НА ИЗМЕНЕНИЕ ПОКАЗАТЕЛЕЙ ТОРГОВО-РАЗВЛЕКАТЕЛЬНОГО ЦЕНТРА

Папаянов Филипп Сергеевич
Главный эксперт – аналитик производственных систем, АО «ТПС Недвижимость»
Аналитик данных, ООО «Мэнпауэр» Московский государственный университет радиотехники, электроники и автоматики (технический университет)
Национальный исследовательский университет "Высшая школа экономики"
Россия, г. Москва
DOI: 10.31618/ESU.2413-9335.2025.1.127.2162

АННОТАЦИЯ

В статье рассматривается метод анализа влияния отдельных арендаторов на изменения показателей торгово-развлекательного центра (ТРЦ). Анализируются существующие подходы к факторному анализу средних значений, выявляет их недостатки, а также предлагается новый метод, основанный на интегральном методе факторного анализа. Описанный метод позволяет корректно учитывать влияние арендаторов на совокупные показатели, устраняя ошибки традиционных методик. Подход применим не только к анализу данных ТРЦ, но и к другим сферам, таким как оценка производительности, затрат на привлечение клиентов и удельных расходов ресурсов. Метод прост в реализации и может быть использован в BI-системах для гибкого анализа данных.

ABSTRACT

The article examines a method for analyzing the impact of individual tenants on changes in the performance indicators of a shopping and entertainment center (SEC). Existing approaches to factor analysis of average values are analyzed, their shortcomings are identified, and a new method based on the integral factor analysis method is proposed. The described method accurately accounts for tenants' influence on aggregate indicators, eliminating errors inherent in traditional methodologies. This approach is applicable not only to SEC data analysis but also to other fields such as performance evaluation, customer acquisition cost assessment, and resource consumption analysis. The method is easy to implement and can be used in BI systems for flexible data analysis.

Ключевые слова: торгово-развлекательный центр, факторный анализ, средний чек, BI-системы, аналитика данных.

Keywords: shopping and entertainment center, factor analysis, average check, BI systems, data analytics.

ВВЕДЕНИЕ

Управляющая компания постоянно анализирует деятельность торгово-развлекательного центра. Качественный анализ позволяет повысить точность планирования, подготовки бюджета, а также принимать взвешенные управленческие решения [1]. В общем случае аналитика сводится к ответам на два вопроса: как изменились показатели всего ТРЦ и как изменились показатели конкретных арендаторов. В число основных показателей входят:

- Арендный доход
- Товарооборот
- Посещаемость
- Количество чеков
- Средний чек
- Доля платежей в товарообороте

При изменении показателя необходимо выделить факторы, за счёт которых это произошло, причём эти факторы могут быть различными: изменение погоды, экономической ситуации, населения города, инфраструктуры (например, строительство дороги, жилого комплекса около ТРЦ) и т. п. Но в ТРЦ любой показатель верхнего уровня складывается из показателей конкретных арендаторов. Например, если выручка ТРЦ выросла на 10% — это не значит, что все арендаторы выросли на 10%, т. к. каждый из них вырос по-разному, а некоторые даже упали.

Для объёмных показателей влияние каждого арендатора вычислить довольно просто (на примере изменения товарооборота к прошлому году):

[Вклад арендатора в изменение ТО] = ([ТО арендатора в новом году] – [ТО арендатора в прошлом году]) / [ТО торгового центра в прошлом году]

Но если речь идёт о производных показателях, то расчёт значительно усложняется. Например, если в ТРЦ снизился средний чек, то это не обязательно означает, что у всех арендаторов он снизился. Возможна ситуация, при которой наоборот, все арендаторы сохранили средний чек на уровне прошлого года (или даже увеличили), но один из арендаторов с низким средним чеком существенно увеличил количество чеков и таким образом «размыл» совокупный показатель. Средний чек вычисляется по формуле *[товарооборот] / [количество чеков]*, формула содержит деление, что не позволяет использовать такой же подход к анализу. Вывести формулу

можно с помощью несложного алгебраического преобразования [2]. Для общего случая средний чек вычисляется по формуле

$$S^{(k)} = \frac{\sum_{i=1}^{n} T_i^{(k)}}{\sum_{i=1}^{n} C_i^{(k)}} \quad (1)$$

где:

$T_i^{(k)}$ - товарооборот i-го арендатора **в k-м периоде**

$C_i^{(k)}$ – количество чеков i-го арендатора **в k-м периоде**

Для задачи факторного анализа сравниваются два периода, поэтому

$k \in \{0, 1\}$; 0 – базовый период, 1 – новый период

n – общее количество арендаторов.

Соответственно, средний чек i-го арендатора в периоде k равен

$$s_i^{(k)} = \frac{T_i^{(k)}}{C_i^{(k)}} \quad (2)$$

Доля i-го арендатора в общем количестве чеков в периоде k есть

$$w_i^{(k)} = \frac{C_i^{(k)}}{\sum_{j=1}^{n} C_j^{(k)}} \quad (3)$$

Совокупный средний чек можно переписать так:

$$S^{(k)} = \sum_{i=1}^{n} w_i^{(k)} s_i^{(k)} \quad (4)$$

Тогда **изменение совокупного среднего чека** с периода 0 на период 1 есть:

$$\Delta S = S^{(1)} - S^{(0)} = \sum_{i=1}^{n} w_i^{(1)} s_i^{(1)} - \sum_{i=1}^{n} w_i^{(0)} s_i^{(0)} = \sum_{i=1}^{n} \left(w_i^{(1)} s_i^{(1)} - w_i^{(0)} s_i^{(0)} \right) \quad (5)$$

Добавим и вычтем дополнительный член с сохранением равенства:

$$\Delta S = \sum_{i=1}^{n} \left(w_i^{(1)} s_i^{(1)} - w_i^{(1)} s_i^{(0)} + w_i^{(1)} s_i^{(0)} - w_i^{(0)} s_i^{(0)} \right) \quad (6)$$

Упростим:

$$\Delta S = \sum_{i=1}^{n} \left(w_i^{(1)} (s_i^{(1)} - s_i^{(0)}) + s_i^{(0)} (w_i^{(1)} - w_i^{(0)} \right) \quad (7)$$

Отсюда естественным образом получается **формула для вклада** i-го арендатора в общее изменение среднего чека:

$$\Delta S_i = w_i^{(1)} (s_i^{(1)} - s_i^{(0)}) + s_i^{(0)} (w_i^{(1)} - w_i^{(0)}) \quad (8)$$

Смысл этой формулы:

1.$w_i^{(1)} (s_i^{(1)} - s_i^{(0)})$ – эффект изменения собственного среднего чека i-го арендатора, взвешенный на его новую долю в общем количестве чеков.

2.$s_i^{(0)} (w_i^{(1)} - w_i^{(0)})$ – эффект изменения доли i-го арендатора в общем количестве чеков, взвешенный на его старый средний чек.

Достоинство такого разложения заключается в том, что все вклады ΔS_i суммируются строго в ΔS. Эту формулу легко вывести, легко реализовать в любом аналитическом приложении (в частности excel), поэтому описанный подход часто используется аналитиками в различных областях, где есть потребность в факторном анализе относительных показателей.

Однако, у метода есть существенный недостаток, который сводит на нет все его преимущества. Рассмотрим таблицу с показателями пяти условных арендаторов:

Показатель	2023			2024			Влияние
	ТО	Чеки	Ср. чек	ТО	Чеки	Ср. чек	
Итого:	**9 000 000**	**56 500**	**159,29**	**10 100 000**	**98 364**	**102,68**	**-56,61**
Арендатор1	500 000	10 000	50,00	100 000	18 364	5,45	-7,83
Арендатор2	1 500 000	1 500	1 000,00	100 000	1 700	58,82	-25,53
Арендатор3	5 000 000	15 000	333,33	5 800 000	17 000	341,18	-29,53
Арендатор4	2 000 000	30 000	66,67	4 000 000	60 000	66,67	5,27
Арендатор5				100 000	1 300	76,92	1,02

У первых трёх арендаторов средний чек снизился, что корректно отражает факторный анализ. Арендатор #4 характеризуется низким средним чеком и самым большим количеством чеков; при этом он вдвое увеличил свои показатели. Легко проверить, что без влияния его изменений средний чек составил бы 118,48, то есть его влияние отрицательное, однако формула говорит об обратном. То же самое касается арендатора #5 – это новый арендатор с низким средним чеком (ниже среднего по выборке), и логично предположить, что его влияние должно быть отрицательным, однако формула даёт положительное значение.

Таким образом, этот подход годится только для приблизительной оценки факторов и в большей степени подходит для наборов данных с низкой волатильностью.

Поэтому часто используется другой подход: для каждого арендатора последовательно моделируется ситуация, при которой его показатели остались неизменными, в то время как показатели остальных арендаторов изменились [3]. Например, для арендатора #1 в 2024 году товарооборот сохранился на уровне 500000, количество чеков осталось 10000, а у остальных арендаторов ничего не поменялось. В этом случае средний чек ТРЦ составил 116,67, что на 13,99 меньше реального (102,68), поэтому влияние арендатора #1 составило -13,99. Таблица, заполненная таким образом, выглядит так:

Показатель	2023			2024			Влияние
	ТО	Чеки	Ср. чек	ТО	Чеки	Ср. чек	
Итого:	**9 000 000**	**56 500**	**159,29**	**10 100 000**	**98 364**	**102,68**	**-38,44**
Арендатор1	500 000	10 000	50,00	100 000	18 364	5,45	-13,99
Арендатор2	1 500 000	1 500	1 000,00	100 000	1 700	58,82	-14,47
Арендатор3	5 000 000	15 000	333,33	5 800 000	17 000	341,18	6,17
Арендатор4	2 000 000	30 000	66,67	4 000 000	60 000	66,67	-15,80
Арендатор5				100 000	1 300	76,92	-0,34

Факторы хорошо описывают влияние, однако их сумма (-38,44) не равна общему отклонению среднего чека (-56,61), что снижает репрезентативность расчёта. Другим недостатком подхода является его трудоёмкость – расчёт приходится выполнять в цикле, и при большом количестве арендаторов это может занять длительное время. Конечно, первую проблему можно решить нормированием, вторую – использованием современных вычислительных средств (например, excel). Но, как правило, арендаторы в отчётах классифицированы по каким-либо признакам (в зависимости от площади, товарной категории и пр.), и при подготовке отчётов нормирование придётся выполнять столько раз, сколько категорий используется для отображения.

Цель данной работы – предложить новый подход к факторному анализу, который будет лишён описанных выше недостатков.

ОСНОВНАЯ ЧАСТЬ

Обозначим искомый показатель как R_i, а также введём вспомогательный показатель a_i:

$$a_i^{(k)} = \frac{T_i^{(k)}}{\sum_{j=1}^{n} C_j^{(k)}} \quad (9)$$

где j = порядковый номер арендатора. Здесь и далее будет использоваться в ситуациях, когда в расчёте по i-му арендатору нужно будет использовать показатели всех арендаторов.

Показатель a_i не имеет самостоятельной ценности, но обладает той же размерностью, что и средний чек, и подчиняется тем же закономерностям: растёт при увеличении товарооборота, падает при увеличении количества чеков. Для простоты обозначим его как «нормированный» чек. Покажем, что сумма нормированных чеков равна среднему чеку ТРЦ:

$$\sum_{i=1}^{n} a_i^{(k)} = \sum_{i=1}^{n} \frac{T_i^{(k)}}{\sum_{j=1}^{n} C_j^{(k)}} = \frac{1}{\sum_{j=1}^{n} C_j^{(k)}} \sum_{i=1}^{n} T_i^{(k)} = \frac{\sum_{i=1}^{n} T_i^{(k)}}{\sum_{j=1}^{n} C_j^{(k)}} = S^{(k)} \quad (10)$$

Следовательно, если разложить на факторы изменение нормированного чека, сумма факторов совпадёт с изменением среднего чека. Для решения задачи воспользуемся классической формулой интегрального метода факторного анализа для модели вида f = x / (y + z) [4]:

$$f = \frac{x}{y+z} \quad (11)$$

$$\Delta_x f = \frac{\Delta x}{\Delta y + \Delta z} \ln \left| \frac{y^{(1)} + z^{(1)}}{y^{(0)} + z^{(0)}} \right| \quad (12)$$

$$\Delta_y f = (\Delta f - \Delta_x f) \frac{\Delta y}{\Delta y + \Delta z} \quad (13)$$

$$\Delta_z f = (\Delta f - \Delta_x f) \frac{\Delta z}{\Delta y + \Delta z} \quad (14)$$

где

•$\Delta_x f$ – изменение значения функции, обусловленное изменением x

•$\Delta_y f$ – изменение значения функции, обусловленное изменением y

•$\Delta_z f$ – изменение значения функции, обусловленное изменением z

Для нормированного чека числитель (x) будет соответствовать товарообороту (T_i), первое слагаемое знаменателя (y) – количеству чеков арендатора (C_i), второе слагаемое знаменателя (z) – количеству чеков всех остальных арендаторов ($L_i = \sum_{j=1}^{n} C_j - C_i$). Для удобства факторы, зависящие от этих аргументов, обозначим одноимёнными индексами: $\Delta_t a_i$, $\Delta_c a_i$ и $\Delta_l a_i$

Для начала определим влияние изменения товарооборота i-го арендатора:

$$\Delta_t a_i = \frac{\Delta T_i}{(\Delta C_i + \Delta L_i)} * \ln\left|\frac{C_i^{(1)} + L_i^{(1)}}{C_i^{(0)} + L_i^{(0)}}\right| \quad (15)$$

Упростим формулу:

$$\Delta_t a_i = \frac{\Delta T_i}{\Delta C} * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| \quad (16)$$

В знаменателе содержится ΔC, поэтому дальнейшие рассуждения справедливы только для ситуаций, когда ΔC не равно 0. Далее определим влияние изменения количества чеков i-го арендатора:

$$\Delta_c a_i = \frac{\Delta a_i - \Delta a(t)_i}{\Delta C} \Delta C_i \quad (17)$$

Формула для третьего фактора (влияние чеков остальных арендаторов) отличается только одним множителем:

$$\Delta_l a_i = \frac{\Delta a_i - \Delta a(t)_i}{\Delta C} \Delta L_i \quad (18)$$

Из этой формулы следует, что изменение количества чеков любого арендаторов оказывает влияние на i-го арендатора. Следовательно, справедливо и обратное: i-й арендатор оказывает влияние на каждого из остальных. Перепишем последнюю формулу:

$$\Delta_l a_i = \frac{\Delta a_i - \Delta_t a_i}{\Delta C} (\Delta C - \Delta C_i) \quad (19)$$

$$\Delta_l a_i = \frac{\Delta a_i - \Delta_t a_i}{\Delta C} * \sum_{j=1}^{n} \Delta C_j - \frac{\Delta a_i - \Delta_t a_i}{\Delta C} \Delta C_i \quad (20)$$

Таким образом, совокупность факторов, связанных с количеством чеков, описывается суммой

$$\Delta_l a_i + \Delta_c a_i = \frac{\Delta a_i - \Delta_t a_i}{\Delta C} * \sum_{j=1}^{n} \Delta C_j \quad (21)$$

В этом равенстве присутствуют индексы i – выбранный арендатор, и j – все арендаторы, включая i. Для наглядности изобразим правую часть этого равенства в виде таблицы:

j / i	Арендатор (1)	Арендатор (2)	…	Арендатор (j)
Арендатор (1)	$(\Delta a_1 - \Delta_t a_1) * \Delta C_1/\Delta C$	$(\Delta a_1 - \Delta_t a_1) * \Delta C_2/\Delta C$		$(\Delta a_1 - \Delta_t a_1) * \Delta C_j/\Delta C$
Арендатор (2)	$(\Delta a_2 - \Delta_t a_2) * \Delta C_1/\Delta C$	$(\Delta a_2 - \Delta_t a_2) * \Delta C_2/\Delta C$		$(\Delta a_2 - \Delta_t a_2) * \Delta C_j/\Delta C$
…				
Арендатор (i)	$(\Delta a_i - \Delta_t a_i) * \Delta C_1/\Delta C$	$(\Delta a_i - \Delta_t a_i) * \Delta C_2/\Delta C$		$(\Delta a_i - \Delta_t a_i) * \Delta C_j/\Delta C$

Каждая ячейка таблицы содержит фактор, показывающий, какое влияние на i-го арендатора оказало изменение чеков j-го арендатора. При i=j это фактор $\Delta_c a_i$. Сумма остальных ячеек строки содержит фактор $\Delta_l a_i$.

В то же время каждый столбец таблицы содержит совокупность факторов влияния j-го арендатора на каждого i-го арендатора. Сумма каждого столбца таблицы содержит совокупное влияние, которое изменение чеков арендатора j оказывает на всех арендаторов, то есть на весь ТРЦ.

Поскольку фактор товарооборота проиндексирован только по i ($\Delta_t a_i$), его влияние зависит только от арендатора i, но не от других арендаторов. Следовательно, размер его влияния на ТРЦ такой же, как на самого арендатора, а искомый показатель R_j будет представлять собой сумму по j-му столбцу плюс фактор $\Delta_t a_j$. Для удобства дальнейшей записи индексы i и j можно поменять местами.

$$R_i = \Delta_t a_i + \Delta C_i * \sum_{j=1}^{n} \frac{\Delta a_j - \Delta_t a_j}{\Delta C} \quad (22)$$

Вынесем ΔC за скобки:

$$R_i = \Delta_t a_i + \frac{\Delta C_i}{\Delta C}\sum_{j=1}^{n}\left(\Delta a_j - \Delta_t a_j\right) \quad (23)$$

Просуммируем Δa_j:

$$R_i = \Delta_t a_i + \frac{\Delta C_i}{\Delta C} * \left(\Delta a - \sum_{j=1}^{n}\left(\Delta_t a_j\right)\right) \quad (24)$$

Подставим $\Delta a(t)_j$:

$$R_i = \Delta_t a_i + \frac{\Delta C_i}{\Delta C} * \left(\Delta a - \sum_{j=1}^{n}\left(\frac{\Delta T_j}{\Delta C} * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right|\right)\right) \quad (25)$$

Выносим за скобки общий множитель:

$$R_i = \Delta_t a_i + \frac{\Delta C_i}{\Delta C} * \left(\Delta a - \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| * \frac{1}{\Delta C}\sum_{j=1}^{n}\left(\Delta T_j\right)\right) \quad (26)$$

Упростим:

$$R_i = \Delta_t a_i + \frac{\Delta C_i}{\Delta C} * \left(\Delta a - \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| * \frac{\Delta T}{\Delta C}\right) \quad (27)$$

Подставим значение $\Delta_t a_i$:

$$R_i = \frac{\Delta T_i}{\Delta C} * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| + \frac{\Delta C_i}{\Delta C} * \left(\Delta a - \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| * \frac{\Delta T}{\Delta C}\right) \quad (28)$$

Раскроем скобки:

$$R_i = \frac{\Delta T_i}{\Delta C} * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| + \frac{\Delta C_i}{\Delta C} * \Delta a - \frac{\Delta C_i}{\Delta C} * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| * \frac{\Delta T}{\Delta C} \quad (29)$$

Вынесем ΔC за скобки:

$$R_i = \frac{\Delta T_i * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| + \Delta C_i * \Delta a - \Delta C_i * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| * \frac{\Delta T}{\Delta C}}{\Delta C} \quad (30)$$

Вынесем логарифм за скобки:

$$R_i = \frac{\left(\Delta T_i - \Delta C_i * \frac{\Delta T}{\Delta C}\right) * \ln\left|\frac{C^{(1)}}{C^{(0)}}\right| + \Delta C_i * \Delta a}{\Delta C} \quad (31)$$

Также значение фактора можно выразить в процентах:

$$R\%_i = \frac{R_i\left(\frac{S^{(1)}}{S^{(0)}} - 1\right)}{\Delta S} \quad (32)$$

Применим эту формулу к тестовым данным:

Показатель	2023			2024			Влияние	Влияние, %
	ТО	Чеки	Ср. чек	ТО	Чеки	Ср. чек		
Итого:	**9 000 000**	**56 500**	**159,29**	**10 100 000**	**98 364**	**102,68**	**-56,61**	**-36%**
Арендатор1	500 000	10 000	50,00	100 000	18 364	5,45	-19,52	-12%
Арендатор2	1 500 000	1 500	1 000,00	100 000	1 700	58,82	-18,88	-12%
Арендатор3	5 000 000	15 000	333,33	5 800 000	17 000	341,18	7,19	5%
Арендатор4	2 000 000	30 000	66,67	4 000 000	60 000	66,67	-24,52	-15%
Арендатор5				100 000	1 300	76,92	-0,89	-1%

Теперь значения фактора отлично характеризуют произошедшие изменения. Например, у арендатора #3 средний чек выше среднего чека ТРЦ, количество чеков увеличилось, поэтому влияние положительное. Новый арендатор #5 с низким средним чеком размыл показатель ТРЦ, его влияние отрицательное.

Отдельно стоит упомянуть, что данный подход корректно работает с группами арендаторов. Представим, что первые два арендатора входят в группу 1, а оставшиеся в группу 2. Тогда значения факторов по группам будут равны сумме факторов по арендаторам, входящим в группу:

Показатель	2023			2024			Влияние	Влияние, %
	ТО	Чеки	Ср. чек	ТО	Чеки	Ср. чек		
Итого:	**9 000 000**	**56 500**	**159,29**	**10 100 000**	**98 364**	**102,68**	**-56,61**	**-36%**
Группа1	2 000 000	11 500	173,91	200 000	20 064	9,97	-38,40	-24%
Группа2	7 000 000	45 000	155,56	9 900 000	78 300	126,44	-18,21	-11%

ЗАКЛЮЧЕНИЕ

Описанный подход позволяет упростить работу по анализу показателей ТРЦ. Но он также применим к любой сфере деятельности, которая требует изучения относительных показателей. Например, формулу можно использовать для оценки изменений:

- производительности труда
- стоимости привлечения клиентов
- времени выполнения заказов
- удельного расхода топлива
- и т. д.

В общем случае предлагаемая методика позволяет анализировать изменения показателей между двумя состояниями системы, поэтому она может быть применена не только для анализа отклонения между двумя периодами, но и при сравнении факта с планом.

Метод прост в реализации, корректно работает с выборками любого размера и с любым набором иерархий, что делает его незаменимым при использовании в BI-системах [5], где пользователь самостоятельно может управлять степенью детализации отчётов.

ЛИТЕРАТУРА

1.Андреев В. Д. Анализ и прогнозирование экономических показателей. – СПб.: Питер, 2019. – 384 с.

2.Шмидт С. Факторный анализ: теория и практика. – М.: Юрайт, 2017. – 312 с.

3.Хилл Т., Вестбрук Р. Маркетинговая стратегия: экономический анализ решений. – СПб.: Питер, 2015. – 480 с.

4.Голубович С. А., Баранов П. В. Интегральный метод факторного анализа в экономических исследованиях // Экономика и анализ данных. – 2018. – №4. – С. 45-59.

5.Новиков Д. А. Бизнес-аналитика: инструменты и технологии обработки данных. – М.: Инфра-М, 2022. – 298 с.

УДК 512

СВОЙСТВА КОЭФФИЦИЕНТОВ РАЗЛОЖЕНИЯ ОТНОШЕНИЯ ПОЛИНОМОВ НА ЭЛЕМЕНТАРНЫЕ ДРОБИ

[1]Перевозников Е.Н., [2]Шахова Е.А.

[1] Канд. физ.-мат. наук, доцент, Военно-космическая академия им. А.Ф. Можайского.

[2] Канд. техн .наук, доцент, Военно-космическая академия им. А.Ф. Можайского.

АННОТАЦИЯ

В работе рассматриваются свойства коэффициентов разложения отношения полиномов на элементарные дроби. Доказана лемма определяющая коэффициенты разложения через корни и коэффициенты полиномов, результаты которой используются для вывода асимптотических выражений релаксационных модулей характеризующих деформационных процессы в полимерных волокнах.

Ключевые слова: свойства коэффициентов разложения отношения полиномов, асимптотические соотношения релаксационных модулей.

1. Разложение отношения полиномов $P_n(x)$, $D_m(x)$ на элементарные дроби

$$\frac{P_n(x)}{D_m(x)} = \frac{1}{C_0}\sum_{i=1}^{m}\frac{A_i}{(x-x_i)} \quad , \quad n \le m+1 \tag{1}$$

используется при вычислении интегралов от дробных функций, при обращении преобразовании Лапласа, при расчете спектров динамических систем [1-3].

2.Покажем, что коэффициенты разложения A_i обладают рядом полезных свойств выраженных следующей леммой

Лемма: если отношение полиномов целочисленных степеней можно представить в виде разложения на элементарные дроби (1), то для коэффициентов разложения A_i имеют место следующие соотношения:

$$a) \quad \sum_{i=1}^{n} A_i = B_0 \quad , \quad b) \quad \sum_{i=1}^{n} A_i / x_i = -C_0 \cdot P_n(0)/D_m(0) \quad ,$$

$$c) \quad A_i = \frac{P_n(x_i)}{\prod_{i\neq\beta}(x_i - x_\beta)} = \left[\frac{C_0(x-x_i)P_n(x)}{D_m(x)}\right]_{x=x_i} . \tag{2}$$

Покажем это: полиномы $P_n(x)$, $D_m(x)$ имеют общий вид

$$P_n(x) = \sum_{k=1}^{n} B_k x^{n-k} \quad , \quad D_m(x) = \sum_{k=1}^{m} C_k x^{m-k} \quad , \tag{3}$$

Если$\{x_i\}$- множество корней полинома D_m (среди которых могут быть и комплексные)

$$D_m(x) = C_0 \prod_{i=1}^{m}(x - x_i) \, , \tag{4}$$

то из (1) для полинома P_n имеем

$$P_n(x) = \sum_{i=1}^{n} A_i \prod_{\beta \neq i}^{m}(x - x_\beta) \, . \tag{5}$$

Из (5) на основании теоремы о единственности полиномиального представления функций, сравнивая (5) и (3) и приравнивая коэффициенты при старшей степени получим первое соотношение из (2) $\sum_{i=1}^{n} A_i = B_0$.

Второе соотношение (б) автоматически следует из (1) при $x=0$.

При $x=x_i$ (и-тому корню полинома D_m) в сумме (5) остается одно слагаемое пропорциональное A_i тогда, с учетом (4), для коэффициента A_i получаем

$$A_i = \frac{P_n(x_i)}{\prod_{\beta \neq i}^{m}(x_i - x_\beta)} = \left[\frac{C_0 \cdot (x - x_i) P_n(x)}{D_m(x)}\right]_{x=x_i} . \tag{6}$$

Т.о. лемма доказана.

3.Используем результаты леммы для получения

асимптотических соотношений для релаксационных модулей характеризующих деформационные процессы в полимерах [3,4].

В [3,4] показано, что процесс релаксации напряжений и ползучести в одномерных полимерных образцах определяется выражениями

$$\sigma(x,t) = E_\infty \varepsilon(x,t) + \int_{-\infty}^{+\infty} dy \int_0^t d\tau \cdot \Delta E_r(x - y, t - \tau) \cdot \partial_\tau \varepsilon(y,\tau) \quad ; \quad (a)$$

$$\varepsilon(x,t) = D_0 \sigma(x,t) + \int_{-\infty}^{+\infty} dy \int_0^t d\tau \cdot \Delta D_r(x - y, t - \tau) \cdot \sigma(y,\tau) \qquad . (b) \tag{7}$$

В (7) σ - механическое напряжение, ε –относительная деформация, ΔE_r

-релаксирующая часть модуля напряжений, ΔD_r - релаксирующая часть модуля ползучести. Для простейших режимов деформирования

$$\varepsilon(x,t) = \varepsilon_0 h(t) h(x) \quad , \quad \sigma(x,t) = \sigma_0 h(t) h(x) \quad ,$$

(где h - функции Хевисайда, ε_0 , σ_0 -начальные значения деформации и напряжения)и усреднения по длине образца-L , релаксационные модули принимают вид [2,3]

$$E_r(t) = E_\infty + \frac{E_0 - E_\infty}{\pi} \int_{-\infty}^{+\infty} \frac{1 - e^{ikL}}{k^2 L} \left[\sum_{\alpha=1}^{3} \chi_\alpha e^{Z_\alpha t}\right] dk \qquad (a) \quad ;$$

$$D_r(t) = D_0 + \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{1 - e^{ikL}}{k^2 L} \left[\sum_{\beta=1}^{3} \frac{\varphi_\beta}{-Z_\beta}\left(1 - e^{Z_\beta t}\right)\right] dk \quad (b) \tag{8}$$

$E_{0,\infty}$, $D_{0,\infty}$ -начальные и равновесные значения модулей, z,k – параметры преобразования Фурье – Лапласа. χ_α , φ_α весовые коэффициенты в (8) являются коэффициентами разложения отношения полиномов $(P/D)_{1,2}$

$$\chi_\alpha = \left\{\frac{P_1(z)(z - z_\alpha)}{D_1(z)}\right\}_{z=z_\alpha} \quad , \quad \varphi_\beta = \left[\frac{P_2(z)(z - z_\beta)}{D_2(z)}\right]_{z=z_\beta} \, , \tag{9}$$

В (9) $D_{1,2}$ спектральные полиномы процессов релаксации и ползучести, $z_{\alpha,\beta}$-соответственно корни спектральных полиномов $D_{1,2}$ [3].

Модули релаксации и ползучести должны удовлетворять асимптотическим соотношениям

$$\lim_{t \to \infty} E_r(t) = E_\infty \qquad (a) \quad , \quad \lim_{t \to 0} E_r(t) = E_0 \quad (b)$$

$$\lim_{t\to\infty} D_r(t) = D_\infty \quad (c) \quad , \quad \lim_{t\to 0} D_r(t) = D_0 \quad (d) \qquad .(10)$$

Действительно: из (8) следует, учитывая , что для установившихся режимов релаксации и ползучести ($Rez_{\alpha,\beta}$ <0) [3] соотношение (10,a) выполняется автоматически. Также автоматически выполняется соотношение (10,d). Соотношения (10,b,c) выполняются при дополнительных условиях

$$\sum_{\alpha=1}^{3} \chi_\alpha = 1 \quad , \quad D_0 \sum_{\beta=1}^{3} \frac{\varphi_\beta}{-z_\beta} = E_\infty^{-1} - E_0^{-1} \, . \qquad (11)$$

Покажем, используя свойства коэффициентов разложения отношения полиномов (2), что условия (11) действительно имеют место . Из (2,а) для коэффициентов χ_α , φ_α см. [3] получим

$$\sum_{\alpha=1}^{3} \chi_\alpha = B_0 = {c_1}^2\theta_1 + c_2^2\theta_2 = 1 \quad , \quad \sum_{\beta=1}^{3} \varphi_\beta = D_0. \qquad . (12)$$

Из (2,б), с учетом явного вида полиномов P_2 , D_2 (см.[2]) следует

$$\sum_{\beta=1}^{3} \left(\frac{\varphi_\beta}{-z_\beta}\right) = \frac{P_2(0)}{D_2(0)} = E_\infty^{-1} - E_0^{-1} = D_\infty - D_0 \quad (13)$$

Важное асимптотическое соотношение для релаксационных модулей.

Дополнительным результатом соотношения (12) является определение модельных параметров – вероятностей распределения потока импульса по молекулярным цепям – каналам (θ_1 , θ_2) , $c_{1,2}$ - скорости переноса импульса по молекулярным цепям.

$$\left.\begin{matrix} c_1^2\theta_1 + c_2^2\theta_2 = 1 \\ \theta_1 + \theta_2 = 1 \end{matrix}\right\} \Rightarrow \theta_1 = \frac{c_2^2 - 1}{c_2^2 - c_1^2} \quad , \quad \theta_2 = \frac{c_1^2 - 1}{c_1^2 - c_2^2} \quad . \qquad .(14)$$

Заключение:

1.В работе рассмотрены свойства коэффициентов разложения отношения полиномов на элементарные дроби, доказана лемма определяющая коэффициенты разложения через корни полинома знаменателя.

2.Выражения коэффициентов разложения применяются для получения асимптотических соотношений релаксационных модулей характеризующих деформационные процессы в одномерных полимерных системах.

3.Коэффициенты разложения также используются для вывода модельных параметров-вероятностей распределения потока импульса по молекулярным цепям-каналам.

Литература

1.Корн Г., Корн Т., Справочник по математике для научных работников и инженеров, М. Наука,1970, (720).

2.Курош А.Г.,Курс высшей алгебры, М. Наука, 1965.

3.Перевозников Е.Н., Скворцов Г.Е.,Динамика возмущений и анализ устойчивости неравновесных систем, СПТЭ, Санкт-Петербург, 2010,(139).

4. Перевозников Е.Н., Степашкина А.С. Неустойчивости деформационных процессов в линейных полимерных системах, Вестник Казанского государственного университета им. А.Н.Туполева, 2022, №2,(32-36).

КОМПЬЮТЕРНЫЕ НАУКИ И ИНФОРМАТИКА

UDC 004.42:336.71

THE EVOLUTION OF IT INFRASTRUCTURE IN BANKS: FROM TRADITIONAL SYSTEMS TO FINTECH

Bolgov S.
Murmansk Arctic University
Murmansk, Russia

ЭВОЛЮЦИЯ ИТ-ИНФРАСТРУКТУРЫ В БАНКАХ: ПЕРЕХОД ОТ ТРАДИЦИОННЫХ СИСТЕМ К ФИНТЕХУ

Болгов С.Н.
Мурманский Арктический Университет
Мурманск, Россия

ABSTRACT

This article examines the evolution of IT infrastructure in the banking sector, focusing on the transition from traditional centralized systems to modern fintech architectures, including microservices and cloud technologies. The key drivers of this transition, such as enhanced flexibility, scalability, and integration with new digital services, are discussed. Special attention is given to security issues, regulatory risks, and challenges related to integrating old and new technologies. The article explores the prospects of using technologies like artificial intelligence, machine learning, blockchain in the future development of banking IT infrastructures. Examples from major banks, such as JPMorgan Chase and Bank of America, are used to illustrate the challenges and benefits of digital transformation.

АННОТАЦИЯ

В статье рассматривается эволюция ИТ-инфраструктуры в банковском секторе, акцентируется внимание на переходе от традиционных централизованных систем к современным финтех-архитектурам, включая микросервисы и облачные технологии. Оцениваются ключевые драйверы этого перехода, такие как повышение гибкости, масштабируемости и интеграция с новыми цифровыми сервисами. Особое внимание уделено проблемам безопасности, регуляторным рискам и вызовам интеграции традиционных и современных технологий. Рассматриваются перспективы использования таких технологий как искусственный интеллект, машинное обучение, блокчейн в контексте будущего развития банковских ИТ-инфраструктур. В статье анализируются примеры крупнейших банков, таких как JPMorgan Chase и Bank of America, которые сталкиваются с вызовами и преимуществами цифровой трансформации.

Keywords: IT infrastructure, fintech, microservices, artificial intelligence, blockchain, digital transformation.

Ключевые слова: ИТ-инфраструктура, финтех, микросервисы, искусственный интеллект, блокчейн, цифровая трансформация.

Introduction

Deep innovation in contemporary banking is driven by the accelerating evolution of fintech technologies, increasing user expectations, and the tightening of regulatory standards. Traditional banking IT infrastructure, based on centralized monolithic systems, does not possess the capacity to meet the demands of flexibility, scalability, and integration of new digital services. In this case, IT infrastructure has been compelled to be rethought by the arrival of novel data management models, process automation, and cybersecurity and thereby allow organizations to resolve modern-day challenges and solidify their competitive standings.

One of the most notable tendencies in banking IT infrastructure is the evolution from monolithic legacy systems towards microservices and modular architectures allowing faster development, reduced operational expenses, and increased reliability of financial services. API-based development is the driving force of this tendency since it guarantees the integration of banking platforms with fintech ecosystems and third-party developers. In addition to this, interest is growing in technology innovations such as artificial intelligence (AI), machine learning (ML), and distributed ledgers that will shape the future of financial services.

This article is focused on the research of evolutionary changes in banks' IT infrastructure, analysis of key drivers of the transition towards fintech models, and consideration of promising areas in the development of banking technologies. System and comparative methods are used in this work for an analysis that will enable one to find patterns of digital transformation in the banking sector.

Main part. Traditional banking IT systems: architecture and limitations

Conventional banking information systems are usually centered on centralized and monolithic architectures: all functions, from transaction processing

to customer data management, are one big system. Systems of that architecture have dominated banking since the very first day of automated banking due to a high level of control and stability combined with relatively small requirements for scalability and flexibility. However, during the last years they have started to face a number of significant limitations that seriously hamper the banks' ability to adapt to rapid market changes.

One of the major **limitations of centralized systems** is that they are not very scalable. As the volume of data grows, so does the load on the infrastructure; traditional systems become increasingly difficult to maintain. Increasing the capacity of such a system requires either the expansion of existing servers, which often leads to a significant increase in equipment and maintenance costs, or the implementation of more complex virtualization and distributed computing mechanisms that are not always justified in terms of profitability [1]. This makes it very difficult for the banks to be responsive to changed user requests, such as an increase in the number of online transactions or the introduction of new financial products.

Moreover, classic systems are rather **inflexible**. It is pretty difficult to make them work with some sort of new technology. For innovative solutions like **AI** or **ML**, one would need to edit software solutions that are already up and running-which in the case of centralized systems often becomes really burdensome and costly. This shows characteristics of the monolithic architecture that these systems used, with strong interdependencies that make it difficult and hazardous to change or update in terms of disturbing other functions of the system. With this view, and as a reason for banks to implement new technology while adapting to new sets of regulatory environments, they are made to carry out resource-intensive and costly reforms. The time that this requires adds to the length and reduces overall competitiveness.

An example of these constraints is the situation with some of the largest USA banks, such as **JPMorgan Chase** in 2023. Their modernization program included multi-year investment in cloud infrastructure, AI, and ML to develop real-time data processing and security. This technology change was aimed at reducing operating costs, enabling faster decision-making, and creating a new customer experience. But this change came at a high immediate cost and operational disruption, particularly during the integration process [2].

Another critical limitation of legacy banking IT systems is **external system integration**. The current operations of banks require interaction with a large number of partners, like fintech institutions, regulators, and other financial institutions. Core systems are inadequately integrated with external systems, and it is hence hard to implement API-centric solutions as well as make the processes slower for interacting with external services. It creates unprecedented regulatory risks and hinders accelerations of innovations such as open banking, which require massive flexibility and compatibility to interact with multiple partners within the ecosystem.

Thus, long-established banking IT infrastructures, while historical and solid, cannot anymore respond to all needs of today's financial markets. They face scalability limitations, an inability to embed new technology, and issues with guaranteeing cooperation with partners. That, in turn, implies moving to more modern architectures, such as microservices and modular systems, with the potential to evolve faster and lower development and maintenance costs.

Technological transition: key drivers and challenges

Banks are moving away from older IT systems to more flexible and scalable setups. The shift is not straightforward and is complex, with compelling reasons and challenges. The most compelling reasons for the move are to improve how well they operate, speed up the launch of new services and products, and collaborate more effectively with external partners in the fast-changing fintech landscape.

Among the major drivers for this technology transition is the **need for scalability**. Classic centralized architectures do not easily scale-up with fast increases in transaction volumes or when the number of users (fig. 1).

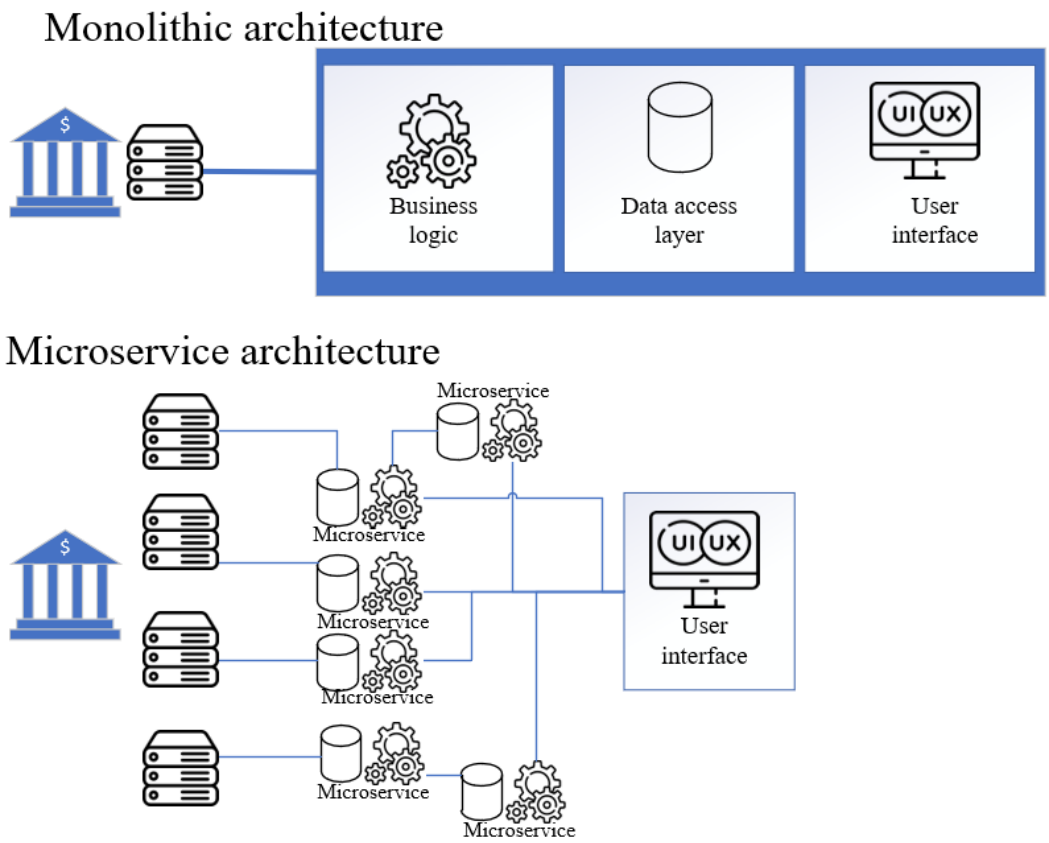


Figure 1. Comparative mapping between monolithic architecture and microservice architecture [3]

Microservice architectures allow banks to flexibly scale individual components of the system without having to rewrite or rebuild the entire infrastructure. This makes banks more flexible and responsive in the face of growing data volumes and user demands.

Another factor in the change is the **bank's need to reduce the time to market for new products**. With legacy systems, making a change or implementing a new function is very labor-intensive and has a long lead time, since the change affects the entire system. In a competitive market where speed is essential, using modular and microservice architecture allows new services to be built and delivered rapidly. Furthermore, applications of technologies like the API-first approach and open banking give rise to more opportunities for interaction with external developers and fintech startups, which lead the innovation processes to accelerate faster [4].

But the technological change is also associated with a number of significant factors. One of them is making new technologies compatible with legacy systems. New solutions must be integrated into existing platforms, and it may become a problem, particularly if legacy systems are monolithic central architectures. Banks are forced to overcome the problem of parallel running of new and old solutions, which is expensive in terms of integration and modernization. This process can also lead to temporary disconnection from customer services, which threatens the reputation and financial position of the bank.

Another major challenge is ensuring security in the context of the transition to more open and connected systems. The implementation of open banking, in which banks make their data available to third-party developers via API, creates new risks to information security [5]. While open interfaces facilitate innovation, they also increase vulnerability to cyberattacks, data leaks, and other security threats. For example, in 2023, a number of large banks in the USA, including **Wells Fargo**, faced data leaks due to unreliable integration with external fintech platforms, which caused public concern and required enhanced data protection measures [6].

Another very big challenge is regulatory compliance. Moving to new technologies and shifting to more flexible architectures will make the banking industry follow international and local data security and protection standards with painstaking care. Considering the current trend of increasing regulation within the fintech sector, such as **GDPR** (General Data Protection Regulation) requirements in Europe or **CCPA** (California Consumer Privacy Act) in the USA, a bank is obliged not only to provide data security but also to follow all the regulations and standards. It means that banks should have not only technical experts but also lawyers who would monitor changes in legislation and adapt technologies to new conditions.

Thus, this transition in technology toward flexible and more innovative architectures of the banks would help them increase their scalability, fasten up their development, and reduce operation costs. Anyway, all the processes above come along with the set of difficulties like integrating old and new technologies, securing the data, and finally reaching all standards prescribed by regulations. This might need an integrated approach, involving substantial investment

both in the development of technological tools and expert people to run those systems.

The future of bank IT infrastructure in the context of fintech

Technological evolution suggests that in the future, the IT infrastructure for banks will be based on modularity, flexibility, and openness principles. The most exciting area would be the development of AI and ML as part and parcel of operations. AI and ML allow automation of data processing processes, predict financial risks, and improve fraudulent act detection. In this context, AI is allowed to analyze huge volumes of real-time transaction data for anomalies that prevent fraud from occurring in the first place [7]. An example of such use is the implementation of ML and AI algorithms in the largest USA banks, such as **Bank of America**, to strengthen security systems and improve customer service [8]. This move allowed Bank of America to reduce fraud investigation-related operational costs and improve customer service through more timely and accurate alerts.

In addition, with the emergence of distributed ledgers such as blockchain, the banks will be capable of changing radically the accounting and interaction with counterparties' strategy. Blockchain technologies can provide transparency and data immutability, which is essential for preserving trust in financial transactions. These technologies guarantee the potential of forgiving transactions, reducing their processing time and lowering the operational expenses cost. Going forward, when the scalability of blockchain is improved, banks will reduce infrastructure expenses drastically and speed up financial transactions that will make finance cheaper and easily accessible to the end users. Another key contributor towards future bank IT infrastructure expansion is the usage of edge computing (table 1).

Table 1.

Technologies and their impact on the future of banking infrastructure

Technology	Characteristics	Impact on IT Infrastructure
AI	Automation of processes, analytics, big data processing.	Increased operational efficiency, cost reduction, improved risk prediction.
Micro-services	Breaking applications into independent components with API.	Enhanced flexibility and scalability of infrastructure.
Blockchain	Distributed ledgers for increasing transaction transparency and security.	Ensures security, reduces intermediary costs, and speeds up transactions.
Cloud technologies	Storing data and running operations on the cloud.	Reduces physical infrastructure costs, increases availability and flexibility.
Edge computing	Processing data is closer to the source of generation.	Reduced latency, improved transaction speed, and enhanced customer experience.

In the long term, edge computing is bound to become inseparable from the future of banking IT infrastructure. Indeed, edge computing enables the processing of data to be performed closer to its origin and source. In particular, real-time work with customer requests and transactions faces critical delays in information processing. Unlike the solutions that centralize data, edge computing minimizes delays and increases the reliability of the systems when there is a high volume and velocity.

The future of banking IT infrastructure will therefore be determined not only by the need to integrate with innovative technologies, but also by the desire to increase the speed and efficiency of operations, as well as to ensure greater security and transparency of financial services. Key aspects will be AI, ML, blockchain, which, combined with modular architectures, will create the foundation for a new generation of banking platforms.

Conclusion

Transition of traditional banking IT systems to fintech architectures forms part of the integral digital transformation of the financial industry. Major challenges include modernizing legacy infrastructures, integrating new technologies such as AI, blockchain, assuring security and regulatory compliance. Yet even with these difficulties, microservice architectures, API-first approach, and open ecosystems create great opportunities for much-needed improvements in terms of agility, scalability, and speed of innovation.

The future of banks will be driven not only by the technological transformation but by the evolution of the role of banks in the financial ecosystem. New technologies will show new ways to improve customer experiences, enhance operational efficiency, and reduce risks. The successful integration of the same in the long run will ensure the sustainable development of the banking sector in the digital era with safer, more accessible, and transparent financial services.

References

1.Kuznetsov I.A., Bobunov A.Yu., Bushuev S.A., Smirnov A.P., Pshichenko D.V. Integration of Big Data into Recommendation Systems: Content Personalization Technologies // Competitiveness in the Global World: Economics, Science, Technology. 2024. №. 9. P. 56-61. EDN: KLLPJU

2.Powering Growth with Curiosity and Heart Annual Report 2023 / JPMorgan Chase // URL: https://www.jpmorganchase.com/content/dam/jpmc/jpmorgan-chase-and-co/investor-relations/documents/annualreport-2023.pdf (date of application: 09.03.2025).

3.Muley Y. Comparative Analysis of Monolithic and Microservices Architectures in Financial Software Development. // J Artif Intell Mach Learn & Data Sci. 2024. №. 2(4). P.1846-1848. DOI: 10.51219/JAIMLD/Yogesh-muley/408

4.Banerjee P. Role of API for the future of Open banking in the USA. 2024.

5.Gupta R. Open banking on the horizon: a scientometric analysis and research agenda. // Electronic Commerce Research. 2024. №. 24(1). P. 577-604. DOI: 10.1007/s10660-023-09722-4 EDN: SSROBR

6.Annual Report 2023 / ELLS FARGO // URL: https://www.wellsfargo.com/assets/pdf/about/investor-relations/annual-reports/2023-annual-report.pdf (date of application: 10.03.2025).

7.Nurdinova K. Integration of Artificial Intelligence Into Accounting as a Tool for Optimization and Risk Management // Bulletin of Science and Practice. 2024. Vol. 10. №. 12. P. 405-410. DOI: 10.33619/2414-2948/109/52 EDN: KMHNQI

8.Annual Report 2023 / Bank of America // URL: https://d1io3yog0oux5.cloudfront.net/_e05654c773adaebb123061b697475523/bankofamerica/db/867/10038/annual_report/BAC+2023+Annual+Report.pdf (date of application: 10.03.2025).

ARCHITECTURAL APPROACHES TO BUILDING HIGHLY AVAILABLE DISTRIBUTED SYSTEMS

Vusatyi Anton Olegovich
Lead Software Engineer - Epam Systems
Türkiye, Istanbul
DOI: 10.31618/ESU.2413-9335.2025.1.127.2163

ABSTRACT

This article analyzes architectural approaches to building highly available distributed systems within the context of modern cloud infrastructures. The relevance of the study stems from the need to ensure uninterrupted service availability and minimize downtime, which is particularly critical given the continuous increase in user requests. The novelty of this work lies in a detailed examination of the integration of microservice patterns, replication methods, and automated cluster management technologies, which enable systems to adapt to peak loads and recover quickly from failures. The study describes mechanisms of active and passive replication, load balancing, and the use of decentralized storage architectures. It also reviews scholarly works dedicated to scaling methods that support high performance and service resilience. Special attention is paid to strategic aspects of microservice architecture that ensure modularity and autonomy of individual components. The aim of the article is to systematize the most effective solutions and demonstrate how their synergy contributes to achieving a high level of availability. To achieve this, the study applies comparative analysis, content analysis of sources, and a historical-analytical review. The conclusion summarizes findings confirming the effectiveness of the presented tools. This article will be of interest to architects, developers, and researchers involved in designing reliable distributed systems.

Keywords: high availability systems, microservice architecture, replication, fault tolerance, scalability, consensus protocols, load balancing, distributed databases, cloud infrastructures, monitoring.

INTRODUCTION

The relevance of this topic is driven by the rapid growth in demand for uninterrupted service availability amid ever-increasing system loads. Both users and organizations now expect systems to withstand peak traffic and consistently deliver high-quality service.

The purpose of this article is to identify and consolidate the key architectural solutions that enable high availability and fault tolerance in distributed systems.

To achieve this goal, the following objectives are addressed: – Analyze the mechanisms for fault tolerance and consensus that ensure consistent data states even in the event of partial failures; – Describe horizontal and vertical scaling techniques, as well as distributed caching and sharding; – Examine microservice patterns and orchestration tools that simplify the management of large clusters and provide flexible service adaptation.

The novelty of this study lies in the systematization of approaches that were previously considered in isolation, and in their practical synthesis for designing highly available architectures.

MATERIALS AND METHODS

The research draws upon scientific and analytical works focused on the design of highly available distributed systems. A. Andhavarapu [1] reviewed modern approaches to fault tolerance in distributed databases, emphasizing multi-leader replication strategies and their impact on response times. N. Badwaik [2] examined architectural solutions that enhance scalability and reliability through horizontal and vertical scaling patterns. P. Chandra [3] focused on architectural principles for achieving a balance between strong consistency and high availability. S. Choudhary and R. Kumar [4] explored service continuity during failures, detailing failover mechanisms and various models of fault-tolerant configurations. F. Dai, M.A. Hossain, and Y. Wang [5] investigated the evolution of parallel and distributed systems, highlighting modern technologies that enable stable performance under high load. V. Iancu and N. Tăpuș [6] proposed decentralized structures based on distributed hash tables, which promote load balancing and high failure readiness. S. Lee and T. Jeong [7] focused on large-scale distributed systems operating in real time and methods for optimizing inter-node network interactions. I. Ortiz [8] examined strategic design aspects of modular microservices, emphasizing their autonomous development and scalability to maintain system availability. S. Saeed and J. Bauer [9] presented mechanisms for building low-latency and fault-tolerant systems in cloud environments, with monitoring and rapid recovery systems playing a central role. Lastly, P.

Ts and D. Devaka [10] provided a detailed case study on implementing scalable and highly available architectures for e-Governance applications, demonstrating the advantages of using distributed clusters in private cloud environments.

A comprehensive set of research methods was used in the preparation of this article:

1.Comparative analysis, which enabled the comparison of various authors' approaches to ensuring high availability, and revealed common trends and differences in the application of replication and load balancing techniques.

2.Content analysis of scientific sources, which facilitated the systematization of key ideas proposed by researchers and the identification of the most relevant tools and solutions for maintaining fault tolerance.

3.Historical and analytical review, covering the evolution of distributed systems and their transition to cloud infrastructures, offering insights into the development of contemporary principles for designing highly available systems.

4.Generalization and synthesis of the obtained data, which allowed for the construction of a comprehensive view on the application of microservice architectures, replication methods, monitoring, and scaling practices necessary to ensure uninterrupted operation of distributed services.

RESULTS

Architectural approaches to building highly available distributed systems rely on principles of scalability, flexibility, and fault tolerance, which are especially critical for today's cloud-based and enterprise infrastructures. In an environment of constantly increasing demand and the need for uninterrupted service delivery, a well-designed architecture becomes the key factor in achieving high availability. Core mechanisms include replication, distributed processing, and automated failover to backup nodes, which collectively help avoid system-wide outages and minimize downtime [1].

A central aspect of such systems is their ability to maintain stable performance under growing user traffic or during abrupt traffic spikes. Research highlights models of horizontal and vertical scaling, as well as combined approaches that leverage both increasing the number of nodes and enhancing the capacity of individual servers or containers [2]. Microservice patterns are widely adopted due to their ability to support the independent development and updating of components, simplifying maintenance and reducing the risk of single points of failure [8]. Each microservice can operate autonomously and be scaled based on specific needs.

The diagram below (see Figure 2) expands a scalable and highly available model through microservices in a clustered arrangement. At the client-facing layer, user traffic reaches a load balancer that distributes incoming requests among active API gateway instances. That gateway operates as a single access point for internal services, handling routing, access control, rate limitations, and observability-focused metrics.

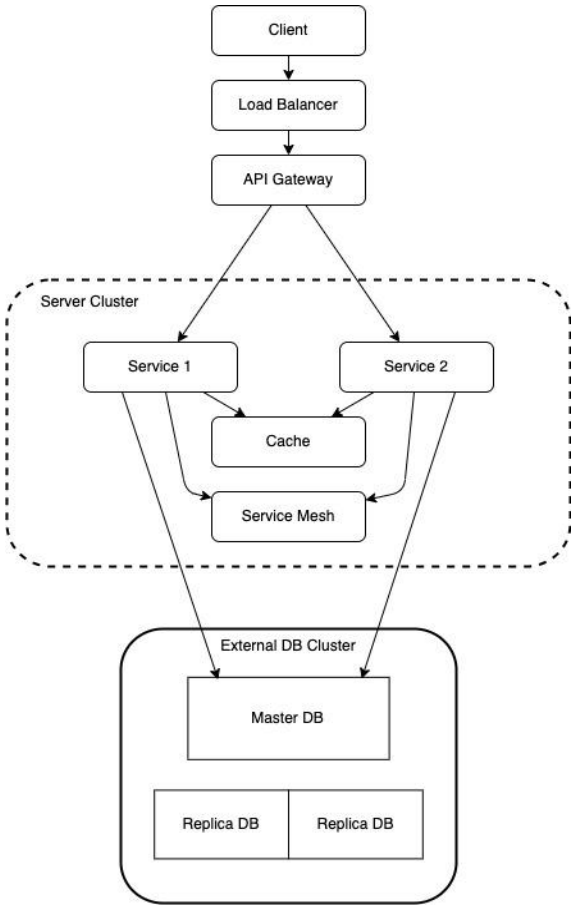


Figure 1 – Microservices-Based High Availability Architecture with Service Mesh and External Database Replication (compiled by the author based on his own research)

Behind that gateway resides a server cluster composed of autonomous microservices, for example Service 1 and Service 2. Each one scales horizontally and exchanges data via efficient RPC protocols supported by a service mesh component. That layer provides resilience enhancements, including circuit breaking, automated retries, and telemetry gathering.

An in-memory cache, such as Redis or Memcached, stands between microservices and the data tier to accelerate performance and relieve stress on the database. Repeatedly accessed details get retrieved directly from that cache, lowering response times and elevating user satisfaction.

Long-term data storage remains in a separate high-availability database cluster, separated at a logical level from the primary compute cluster. The database system follows a primary-replica configuration, with a master node assigned for write operations and replicas handling read queries while acting as fallback nodes in case of disruptions. Such a layout strengthens fault isolation, improves data durability, and allows independent management of database infrastructure without interfering with application deployments.

If a node fails or becomes unreachable, the SLB stops routing traffic to it and automatically redirects requests to a healthy replacement. As demand increases, additional servers can be added to the group, and the SLB will begin utilizing the new resources without requiring reconfiguration. When session persistence (sticky sessions) is necessary, the service ensures that subsequent requests from the same user are directed to the same server. This approach facilitates balanced traffic distribution and improves overall system fault tolerance [10].

To enhance failure resilience, various data replication strategies are employed, including multi-leader schemes where multiple nodes can accept writes simultaneously and then synchronize their change logs [1]. This reduces response times and increases overall system resilience, as the failure of one replica can be quickly mitigated by switching to another [4]. A key enabler of this process is the consensus protocol, which ensures data consistency and allows the system to maintain a synchronized state across replicas despite delays or node failures [3].

Stream processing plays a vital role by enabling real-time data handling and immediate detection of potential anomalies (e.g., overloads or failures) as they begin to emerge [9]. Event-driven architecture is often used in this context, where key operations are asynchronously propagated throughout the system using pub/sub mechanisms [7]. This increases the flexibility of distributed solutions and eliminates the need for synchronous blocking in high-volume data scenarios.

To handle high loads while maintaining low latency, sharding and distributed caching are widely adopted. Sharding involves dividing a database into logical segments (shards) and distributing them across different nodes, which enables even load distribution and efficient data volume management [6]. Caching systems reduce redundant access to primary storage, accelerate access to frequently used data, and serve as an additional buffer layer between applications and the database [5].

An equally critical factor is the organizational and technical preparedness for failures. According to several studies, a high level of resilience is achieved through regular stress testing, continuous monitoring of system metrics, and the implementation of automatic recovery mechanisms that respond immediately upon detection of failures [3]. Modern solutions increasingly rely on container platforms and orchestration tools (such as Kubernetes and Docker Swarm) for centralized cluster management, automated scaling, and simplified deployment of microservices [8]. Closely tied to this are configuration management systems (e.g., Ansible, Terraform), which help minimize human error and ensure consistent settings wherever needed [4].

Network interaction also requires particular attention. In highly available architectures, load balancers play a crucial role by distributing incoming requests across multiple nodes and monitoring their health in order to quickly remove any faulty node from rotation [10]. In geographically distributed infrastructures, low-latency topologies and adaptive routing algorithms are increasingly used to take into account current network conditions and user locations, enabling optimal traffic distribution [7].

To better illustrate the practical approaches used in the development of highly available distributed systems, two summary tables below describe specific patterns, technologies, and scaling strategies. The information has been synthesized from a number of sources, particularly [2], [4], [7], [8], and [9].

Table 1

Patterns and Mechanisms for Building a Highly Available Distributed Architecture (Source: synthesized by the author based on [4; 9])

Pattern/Mechanism	Description	Advantages	Possible Limitations
Active-active replication	Multiple nodes handle requests concurrently and synchronize data in the background.	- High availability- Instant failover- Minimal downtime	- Complex configuration- Requires additional network resources
Failover controller	Redirects traffic and processes to backup resources upon failure of a primary node or region.	- Automated recovery- Reduced downtime	- Requires precise monitoring setup- Risks with failover logic
Multi-master replication	Each node can perform write operations and reach consensus across nodes.	- Faster write operations- High fault tolerance	- Complex consensus algorithms- Potential version conflicts
Load balancing	Distributes incoming traffic across cluster nodes for even workload.	- Increased throughput- Reduced latency- Flexible scalability	- Requires extra infrastructure- Continuous health checks needed
Automated monitoring	Real-time system metric analysis using tools (e.g., Prometheus, Zabbix).	- Early issue detection- Faster response- Supports bottleneck prediction	- Possible false positives- Requires continuous oversight

Table 2, titled “Scaling and performance maintenance techniques”, draws on developments and cases from [2], [7], and [8]. It provides a brief overview of strategies that allow distributed systems to maintain stable response times and reliability under increasing load.

Table 2

Scaling and Performance Maintenance Methods (Source: synthesized by the author based on [2; 7; 8])

Method	Brief Description	Pros	Cons / Implementation Considerations
Vertical scaling (Scale Up)	Increasing the hardware resources of a single machine (CPU, RAM, SSD)	- Simple implementation- No need to rework architecture	- Hardware limits- Potentially costly
Horizontal scaling (Scale Out)	Adding more servers/containers and distributing load among them	- High flexibility- Linear performance growth potential	- More complex networking and management- Requires balancing
Sharding	Dividing a database into logical segments (shards) on different nodes	- Load reduction on individual nodes- Enables independent scaling	- Complex query logic- Needs a clear sharding and replication strategy
Caching (Distributed Caching)	Use of in-memory distributed stores (e.g., Redis, Memcached)	- Reduced access latency- Offloads primary database- Faster response	- Requires a solid cache invalidation strategy- Potential RAM overload
Automatic orchestration	Centralized container cluster management and scaling (e.g., Kubernetes)	- Simplified deployment- Resource isolation- Responsive to demand	- Requires orchestration expertise- Adds abstraction and complexity

Together, these tables reinforce the broader concept of building highly available systems, emphasizing the importance of well-structured replication, fault-tolerant protocols, and efficient scaling. The described patterns and methods make it possible to achieve high availability metrics while maintaining performance during spikes in traffic or partial component failures.

In a related study the authors conducted a quantitative evaluation using two linear graphs to assess operational performance under failure conditions [12]. The first graph (see. Figure 1) measures the percentage of time systems remained operational under various failure scenarios. Data indicate that for node failures, active-active configurations achieved 99.99% uptime, while active-

passive setups, synchronous replication, and asynchronous replication reached 99.85%, 99.95%, and 99.90% respectively. Under network partitions, active-active configurations recorded 99.80% uptime; asynchronous replication outperformed both active-passive (99.75%) and synchronous replication (99.70%). Similarly, during database failures and sudden traffic surges, the measured uptimes consistently favored active-active configurations.

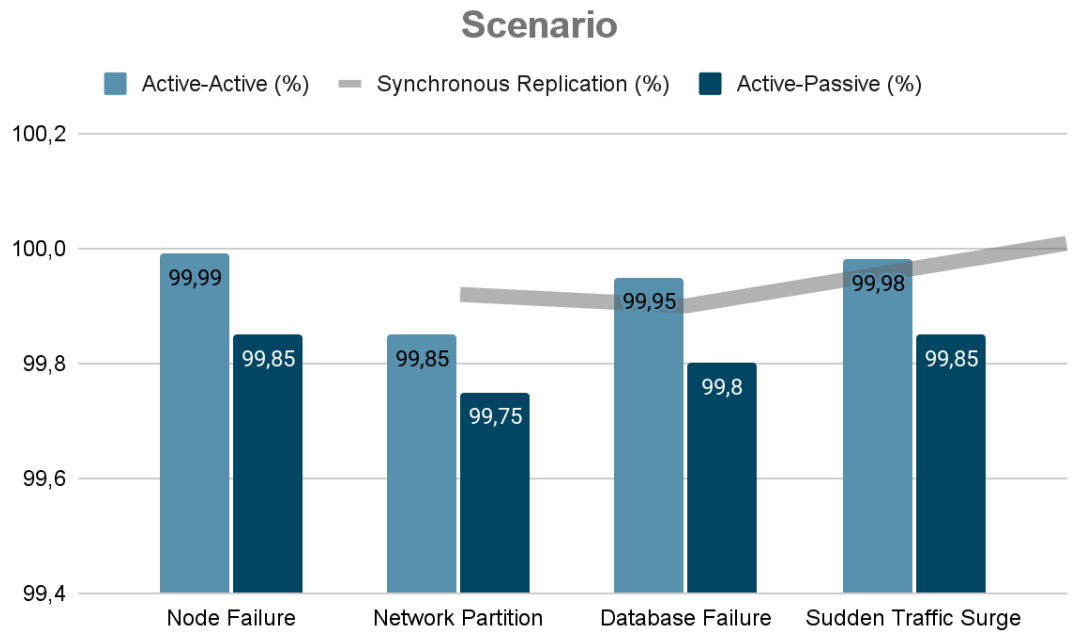


Figure 1. Uptime Percentage Analysis [12]

The second graph presents the number of successful requests processed per second across comparable scenarios. Under node failure conditions, active-active configurations processed approximately 2000 requests per second, outperforming active-passive (1800 req/s), synchronous replication (1900 req/s), and asynchronous replication (1950 req/s). During sudden traffic surges, the active-active setup reached 2100 req/s, significantly exceeding the throughput of active-passive (1900 req/s), synchronous replication (2000 req/s), and asynchronous replication (2050 req/s).

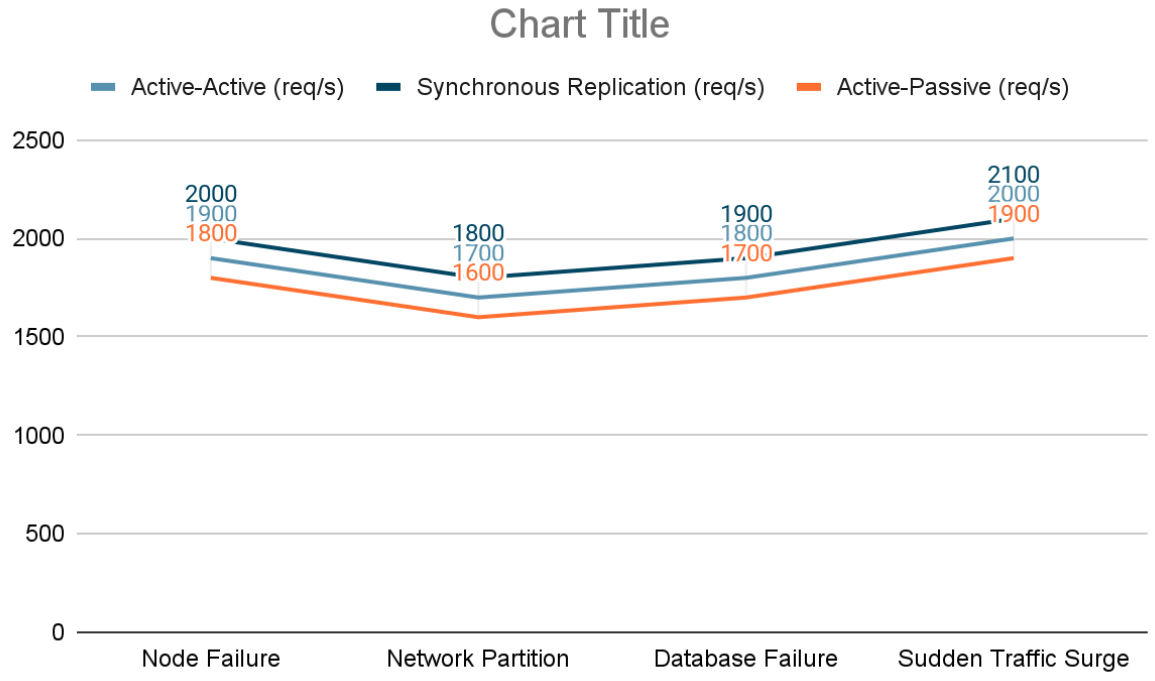


Figure 2. System Throughput [12]

These findings substantiate the superior performance and recovery efficiency of active-active configurations in high-demand environments, providing objective evidence for improved fault tolerance and responsiveness in distributed systems.

In conclusion, modern architectural approaches to highly available distributed systems represent a combination of technological and organizational strategies: a microservices model, distributed data storage and processing, proactive monitoring mechanisms, and well-planned replication and scaling techniques. The implementation of multi-layered fault-tolerance strategies ensures that systems can continue operating even after the failure of certain components. Moreover, the adoption of event-driven architecture makes systems more adaptable and responsive to external changes [9]. All these methods are integrated within orchestration and configuration management environments, which guarantee state consistency and optimal resource utilization [2]. As a result, a highly available distributed system provides a reliable foundation for businesses capable of meeting the demands of the modern market and sustaining high operational loads without compromising service quality.

DISCUSSION

The analysis of the reviewed studies demonstrates that the choice of specific technological tools—such as orchestrators, containers, or service meshes—is not always the decisive factor for the success of a highly available distributed system. A critical role is played by the flexibility of the development team's organizational structure and the ability to implement changes quickly without major downtime [3; 8]. Current practice shows that microservice decomposition, when combined with well-designed replication and load balancing strategies, enables rapid scaling of narrowly focused services without increasing latency for the rest of the system [2; 5]. At the same time, any failures occurring in individual microservices are localized within their environment, reducing the risk of a cascading failure effect [4].

However, studies [6; 7] emphasize that the widespread adoption of microservice architecture also introduces significant complexity in configuration and debugging. For example, the presence of numerous independent databases and asynchronous communication channels increases the risk of data inconsistency, requiring the implementation of consensus algorithms and adaptive sharding strategies. Monitoring also becomes more challenging: a multitude of microservices generates a heterogeneous stream of logs and metrics that must be analyzed in real time to detect failures promptly [8; 9]. These costs, however, are offset by the ability to optimize and release exactly those services under the heaviest load—whether it's machine learning modules or business logic components [1; 10].

Another important aspect is the distribution of responsibility within the development team. A service-oriented approach implies that each team manages the full lifecycle of one or more microservices, including the choice of technologies and testing [2; 4]. While this autonomy speeds up processes, it also demands unified policies for security and metrics collection; otherwise, the uncontrolled proliferation of services can lead to redundant functionality and infrastructure fragmentation [8]. Various authors [7; 5] agree that special attention must be given to the protection of API gateways and event channels, as they are the primary points of entry into the system.

There is broad consensus that CI/CD processes (continuous integration and delivery), combined with containerization tools (such as Docker and Kubernetes), greatly simplify update rollouts and help maintain high availability [3; 9]. Proper build and deployment configuration, complemented by centralized monitoring and configuration management systems, ensures that failures in individual nodes do not affect the overall health of the application [1; 6]. At the same time, human factors and organizational aspects become increasingly important: teams operating under Agile or DevOps paradigms are better positioned to respond to issues quickly and maintain a unified architectural direction [9; 10].

In summary, the accumulated experience confirms that implementing microservice architecture in high-load and dynamically evolving systems yields significant benefits—provided that coordination, standardization, and the use of modern orchestration, containerization, and monitoring tools are well established [7; 8]. Beyond the technological components, success also hinges on sound organizational and managerial practices that enable effective distribution of responsibility and quality control of microservice components.

CONCLUSION

The conducted analysis confirms the effectiveness of combining microservice architecture, replication mechanisms, and orchestration tools to achieve high availability in distributed systems.

The first objective, focused on fault-tolerance mechanisms, demonstrated that implementing both active and passive replication, along with consensus algorithms, significantly reduces the risk of systemic failures. The second objective, which addressed scaling and distributed caching, revealed a substantial performance increase when load balancing is properly configured and data is partitioned into shards. The third objective, involving the study of microservice patterns, confirmed their ability to enhance system flexibility and simplify service updates without compromising availability.

Thus, the article's goal—to identify and consolidate key architectural solutions necessary for ensuring the continuous operation of distributed systems—has been achieved. The conclusions presented may serve as a foundation for designing scalable and reliable systems of various sizes.

REFERENCES

1. Andhavarapu A. Building Resilient Distributed Databases: Modern Approaches to High Availability. – 2025. – February. – URL: https://www.researchgate.net/publication/389534235_Building_Resilient_Distributed_Databases_Modern_

Approaches_to_High_Availability (accessed: April 3, 2025).

2. Badwaik N. Designing System Architecture with High Availability and Scalability // American Research Journal of Computer Science and Information Technology. – 2024. – Vol. 7, No. 1. – P. 6–10. – DOI: 10.21694/2572-2921.24002.

3. Chandra P. Reliability-Driven Architecture Design for Distributed Systems: Key Principles and Practical Approaches // International Journal of Research in Computer Applications and Information Technology (IJRCAIT). – 2025. – Vol. 8, No. 1. – P. 2583–2597. – DOI: https://doi.org/10.34218/IJRCAIT_08_01_187. – URL: https://iaeme.com/MasterAdmin/Journal_uploads/IJRCAIT/VOLUME_8_ISSUE_1/IJRCAIT_08_01_187.pdf (accessed: April 3, 2025).

4. Choudhary Rajesh S., Kumar R. High Availability Strategies in Distributed Systems: A Practical Guide. – 2025. – January. – URL: https://www.researchgate.net/publication/388075854_High_Availability_Strategies_in_Distributed_Systems_A_Practical_Guide (accessed: April 3, 2025).

5. Dai F., Hossain M. A., Wang Y. State of the Art in Parallel and Distributed Systems: Emerging Trends and Challenges // Electronics. – 2025. – Vol. 14. – Article ID: 677. – DOI: 10.3390/electronics14040677. – URL: https://doi.org/10.3390/electronics14040677 (accessed: April 3, 2025).

6. Iancu V., Ţăpuş N. Towards a Highly Available Model for Processing Service Requests Based on Distributed Hash Tables // Mathematics. – 2022. – Vol. 10. – Article ID: 831. – DOI: 10.3390/math10050831. – URL: https://doi.org/10.3390/math10050831 (accessed: April 3, 2025).

7. Lee S., Jeong T. Large-Scale Distributed System and Design Methodology for Real-Time Cluster Services and Environments // Electronics. – 2022. – Vol. 11. – Article ID: 4037. – DOI: 10.3390/electronics11234037. – URL: https://doi.org/10.3390/electronics11234037 (accessed: April 3, 2025).

8. Ortiz I. Strategic Approaches to Building Highly Scalable, Modular, and Fault-Tolerant Microservices: Enhancing Application Development, Deployment Efficiency, and Long-Term Maintainability in Modern Distributed Systems // International Journal of Social Research. – 2023. – July. – URL: https://www.researchgate.net/publication/386222481_Strategic_Approaches_to_Building_Highly_Scalable_Modular_and_Fault-Tolerant_Microservices_Enhancing_Application_Development_Deployment_Efficiency_and_Long-Term_Maintainability_in_Modern_Distributed_Sy (accessed: April 3, 2025).

9. Saeed S., Bauer J. Building Resilient Distributed Systems for High Availability and Low Latency in Cloud Environments // Education and Computing. – 2025. – March. – URL: https://www.researchgate.net/publication/389986975_Building_Resilient_Distributed_Systems_for_High_Availability_and_Low_Latency_in_Cloud_Environments (accessed: April 3, 2025).

10. Ts P., Devaka D. A Scalable and Highly Available Distributed Architecture for e-Governance Applications on Private Cloud Platform // International Journal of Computer Sciences and Engineering. – 2019. – Vol. 7, No. 3. – P. 811–814. – DOI: 10.26438/ijcse/v7i3.811814. – URL: https://www.researchgate.net/publication/335803639_A_Scalable_and_Highly_Available_Distributed_Architecture_for_e-Governance_Applications_on_Private_Cloud_Platform (accessed: April 3, 2025).

11. Palachi E. What is distributed architecture? Know the types and key elements.– 2024. – September. – URL: https://vfunction.com/blog/distributed-architecture/ (accessed: April 3, 2025).

12. Siddharth Ch., Kumar R. High Availability Strategies in Distributed Systems: A Practical Guide // International Journal of Research in all Subjects in Multi Languages – 2025. – Vol. 13, No. 01. – P. 110-130.

UDC 004.056.5

ZERO TRUST MODELS IN WEB DEVELOPMENT

R. Garifullin
Saint Petersburg Electrotechnical University «LETI»
Professora Popova str., 5, Saint Petersburg, Russia, 197022

МОДЕЛИ НУЛЕВОГО ДОВЕРИЯ (ZERO TRUST) В ВЕБ-РАЗРАБОТКЕ

Гарифуллин Р.Ш.
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»
им. В.И. Ульянова (Ленина)
ул. Профессора Попова, 5, Санкт-Петербург, Россия, 197022

ABSTRACT

This article examines the Zero Trust model as a key approach to ensuring security in web development. Methods for integrating it into the architecture of web applications are studied, including multi-factor authentication, context-based authorization, micro-segmentation, and continuous interaction monitoring. The model's impact on preventing data breaches, enhancing resilience to potential cyberattacks, and improving overall system security is analyzed. Practical recommendations for adapting the Zero Trust concept to modern web development requirements and effective risk management are presented.

АННОТАЦИЯ

В данной статье рассматривается модель нулевого доверия (Zero Trust) как ключевой подход к обеспечению безопасности в веб-разработке. Изучаются методы ее интеграции в архитектуру веб-приложений, включая многофакторную аутентификацию, авторизацию на основе контекста, микросегментацию и постоянный мониторинг взаимодействий. Исследуется влияние модели на предотвращение утечек данных, повышение устойчивости к возможным кибератакам и улучшение общей защищенности систем. Даются практические рекомендации по адаптации концепции Zero Trust для современных требований веб-разработки и эффективного управления рисками.

Keywords: zero trust, web development, data security, data breaches, micro-segmentation, authentication, authorization, monitoring, cybersecurity threats.

Ключевые слова: нулевое доверие, веб-разработка, безопасность данных, утечка данных, микросегментация, аутентификация, авторизация, мониторинг, киберугрозы.

Introduction

Modern web development is faced with such an unprecedented rise in vulnerabilities related to data breaches, hacking, and unauthorized access. In the age of digitalization, when web applications are turning into the focal point of interaction between business and user, traditional ways of guaranteeing security to information are quite insufficient. This demands the implementation of modern strategies that can ensure strong data protection with minimal risk. Examples include the Zero Trust models that operate on the principle of complete distrust in any system, user, or process unless it has been verified.

The utilization of Zero Trust in web application architecture provides significant security enhancements through the requirement of strict authentication, approval, and observance of all system interactions. The concept has effectively been employed in the development of mission-critical and high-load applications, effectively repelling data breaches and security attacks. Adopting this model requires a thorough understanding of its principles and the development of practical recommendations tailored to the specific needs of web development. The purpose of this study is to explore the potential for integrating Zero Trust principles into web application architectures and evaluate their impact on improving security and reducing the risk of data breaches.

Main part. Principles and foundations of the Zero Trust model

The concept of Zero Trust was brought to light due to the emergence of highly complex, state-of-the-art threats, for which previously designed security models were inefficient. Traditional information security has been built on a perimeter protection concept, assuming threats are outside the network and users and devices inside are implicitly trusted. All this has proved to be vulnerable in front of emerging cyber threats, whereby internal risks and sophisticated attacks, like phishing and credential theft, have become rampant [1].

Zero Trust is founded on the principle of complete distrust toward all system elements, whether users, devices, or processes, regardless of their location. This approach requires access to any resource to be granted only after strict verification of identity, device, and context, as well as compliance with all security policies. A central component of this model is the continuous verification of all participants in the system, including repeated checks with every access request. According to 2024 statistics [2], only about a quarter of professionals within companies have already implemented the Zero Trust model, while approximately one-third are in the process of adopting this technology (fig. 1).

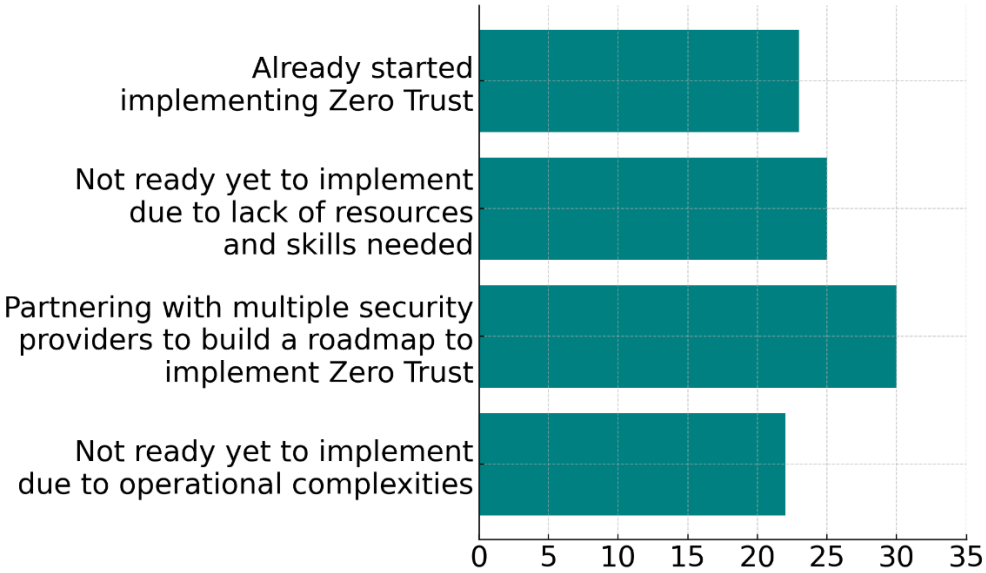


Figure 1. Survey of Zero Trust implementation plans worldwide 2024, %

The foundational principles of Zero Trust encompass several demanding aspects. First is that the model assumes that the security perimeter is no longer confined to physical or network infrastructure. In the period of cloud technologies, remote work, and decentralized systems, access can originate from anywhere in the world, making verification at every stage essential. Second is the principle of «least privilege», that plays a pivotal role in maintaining security. Access to data and resources is granted only to the extent necessary to perform a specific task, thereby reducing the likelihood of unauthorized use.

The Zero Trust model also incorporates security management through micro-segmentation. Unlike traditional approaches that secure an entire network as a single entity, Zero Trust divides the infrastructure into smaller segments, each governed by its own access policies. This strategy minimizes the impact of a compromised system component, as an attacker cannot gain access to the entire network [3].

The advantages of the Zero Trust model are evident. It substantially diminishes the risk of data breaches and mitigates the impact of cyberattacks through multi-layered verification and granular access control. Continuous monitoring of user and device activity enables the detection of anomalies and rapid response to potential incidents. These attributes make Zero Trust very relevant for web development, where safeguarding user and client data is one of the top priorities [4].

The Zero Trust concept serves as a powerful tool for ensuring security in today’s digital environment. Its principles, including multi-layered verification, least privilege enforcement, and micro-segmentation, provide a robust foundation for building secure systems. Its effective implementation also requires a thorough understanding of both the theoretical underpinnings and the practical mechanisms of execution.

Integration of Zero Trust into web application architecture

The integration of the Zero Trust model into web application architecture necessitates a reevaluation of traditional approaches to design and security. Unlike conventional solutions focused on perimeter-based network protection, Zero Trust is embedded directly into the web application structure, encompassing access management processes, component interactions, and data protection across all levels.

A key aspect of integration involves building an architecture grounded in principles of isolation and verification. Web applications often comprise numerous components, including the server-side infrastructure, user interface, databases, and application programming interface (API) [5]. Zero Trust mandates that each of these components interacts with others through identity, context, and permission verification. Every user request to a database must be checked for access rights, even if the user has previously been authenticated. This approach inhibits credential-compromise or weakness-based attacks in individual system components.

Access control within the Zero Trust paradigm also entails the utilization of advanced authentication and authorization methods. Multi-factor authentication (MFA) is among the most effective solutions, which offers additional layers of verification such as biometrics, one-time passwords (OTP), or tokens. Authorization needs to be context-aware access models, where a user's account is evaluated on the basis of permissions and the current context, such as location, device, or request time.

Zero Trust implementation in web development necessitates the active use of micro-frontend technologies and API gateways. Micro-frontends refer to small, individual interface modules that enable application functionality to be separated and the impact of potential attacks to be reduced. In the event of an attack on a module, it is not able to harm others. API gateways, on the other hand, serve as centralized nodes that manage access to all API and enforce Zero Trust

policies, including data encryption and activity monitoring.

Special attention is given to monitoring and analyzing user behaviour. Zero Trust requires infrastructures that will be able to monitor device and user behavior in real-time. These systems have behavior analytics software that is capable of detecting anomalous behavior, which may constitute a threat like suspicious search terms or out-of-pattern behavior inside the system.

One of the most significant aspects of integration is encrypting data at all levels of interaction. Web applications that are developed based on Zero Trust principles utilize protocols such as TLS (Transport Layer Security) to provide secure data transmission and prevent interception by attackers. Data encryption methods are applied to stored information, ensuring its security even in the event of server compromise.

Effective integration of Zero Trust into web application architecture requires meticulous configuration of security policies. These policies must be both detailed and flexible to align with the dynamic nature of web applications and their users. Automation tools are employed to adapt access rules to changing conditions in real-time, increasing responsiveness and security.

As a consequence, web application architectures built on Zero Trust principles become more resilient to threats while providing a high level of data security. This approach minimizes the risks of data breaches and attacks while maintaining performance and usability. Implementing Zero Trust represents an important step for developers aiming to create modern, secure, and reliable web solutions.

Effectiveness of Zero Trust in enhancing security and mitigating risks

The actual success of the Zero Trust model in web development is all about its dynamic adaptation capability to the upcoming threats and assurance of maximum data security. In such a context, when traditional protection methods are not good enough, the implementation of Zero Trust reduces the likelihood of an attack and its potential impact.

Probably the most important result of working with Zero Trust is that it diminishes data breach risks. Providing severe authentication and authorization policies based on contextual factors, it does not allow attackers to use account credentials that have been stolen. If the login and password of a certain user were compromised, access will be blocked in the case when the request comes from an unverified device or suspicious location. It decreases the possibility of breaches and, therefore, protects sensitive information. According to researches [6], as companies increasingly adopt Zero Trust in their workflows, they achieve significant results in reducing the cost of data breaches and the time required to detect and respond to potential threats (table 1).

Table 1.

The impact of Zero Trust security model adoption on data breach costs and incident detection time

Year	Organizations adopting Zero Trust, %	Average cost of data breach, million dollars	Average time to detect and respond to security incidents, days
2018	15	3.86	120
2019	22	3.92	105
2020	35	4.12	90
2021	50	4.24	75
2022	60	3.98	60
2023	75	3.65	45

Zero Trust minimizes the impact of attacks through micro-segmentation and access control as well. Isolation of the system into separated segments prevents the propagation of threats across the infrastructure. Even if an attack occurs on one segment, the attacker will not have any access to other system assets. This is particularly critical for complex web applications, where an insecurity attack in a section may potentially render the entire system vulnerable.

The implementation of Zero Trust enhances resilience to modern threats, such as advanced attacks such as Advanced Persistent Threats (APT). They are aimed at unauthorized access to systems and can last for many months. The real-time authentication and activity audit that define the Zero Trust model also make such attacks less straightforward to carry out because anomalies in user or device action are detected and neutralized early on.

Another advantage of the model is that it can help increase system transparency. By continuously tracking user interactions and application components, developers and administrators have a complete view of the current security state. This simplifies threat detection and assists in exposing vulnerabilities within the system so that they may be patched before any malicious party can utilize them.

Zero Trust also increases user trust in web applications. As data breaches become increasingly common, clients place greater value on companies that prioritize the protection of their information [7]. The implementation of Zero Trust demonstrates a serious commitment to security and reduces the likelihood of reputational damage associated with such incidents.

Despite its numerous advantages, the adoption of Zero Trust may encounter certain constraints. Strict security policies can affect the usability of web applications if they are not configured with sufficient flexibility. To mitigate this, it can be essential to implement Zero Trust gradually, starting with the most vulnerable areas, and to adapt security policies to align with user needs [8].

The effectiveness of Zero Trust is witnessed by its ability to prevent attacks and data breaches and mitigate their effects. Through the application of new techniques

in authentication, micro-segmentation, and real-time monitoring, the model is an invaluable resource for web application security.

Conclusion

The Zero Trust model is an innovative approach to web application security, gaining increasing relevance in the wake of increased cyber threats. Leveraging concepts such as multi-layer verification, micro-segmentation, and least privilege policy, Zero Trust reduces the attack surface for data breach, contains the impact of an attack, and establishes user trust. Implementation of this model demands a reconsideration of current design and system management practices along with the use of advanced authentication, authorization, and monitoring technologies. Despite potential challenges in adaptation, Zero Trust has proven its effectiveness as a versatile solution for building secure, reliable, and user-friendly web applications, making it an integral component of development strategies in the era of digital transformation.

References

1.Yeoh W. Zero trust cybersecurity: Critical success factors and A maturity assessment framework // Computers & Security. 2023. Vol. 133. P. 103412. DOI: 10.1016/j.cose.2023.103412 EDN: VGRGCA

2.How are you planning to implement Zero Trust across your extended environment? / Statista // URL: https://www.statista.com/statistics/1458903/global-zero-trust-implementation-plans/ (date of application:12.04.2025).

3.Basta N. Towards a zero-trust micro-segmentation network security strategy: an evaluation framework // NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium. 2022. P. 1-7. DOI: 10.1109/NOMS54207.2022.9789888

4.Aluev A. Addressing security issues in Node.js applications: the economic implications of increased security // International Journal of Humanities and Natural Sciences. 2024. Vol. 9-1(96). P. 94-98.

5.Qazi F. A. Study of zero trust architecture for applications and network security //2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET). 2022. P. 111-116. DOI: 10.1109/HONET56683.2022.10019186

6.Gudimetla, S. Zero Trust security model: implementation strategies and effectiveness analysis. International Research Journal of Innovations in Engineering and Technology. 2024. Vol. 11. P. 1186-1194.

7.Kuznetsov I.A. Forecasting and analysis of user behavior in mobile applications // Trends in the development of science and education. 2024. №. 107 (8). P. 165-168. DOI: 10.18411/trnio-03-2024-437. EDN: IQUSMS

8.Buck C. Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero-trust // Computers & Security.2021. Vol. 110. P. 102436. DOI: 10.1016/j.cose.2021.102436 EDN: DHDZVA

ЭНЕРГОЭФФЕКТИВНОСТЬ В РАЗРАБОТКЕ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ: АЛГОРИТМЫ И СТРАТЕГИИ

Телегин Валентин Александрович
Технический директор департамента мобильной разработки ООО «Ростелеком Информационные Технологии»
Москва, Россия

ENERGY EFFICIENCY IN MOBILE APPLICATION DEVELOPMENT: ALGORITHMS AND STRATEGIES

DOI: 10.31618/ESU.2413-9335.2025.1.127.2164

АННОТАЦИЯ

В статье проведён анализ проблем энергоэффективности в разработке мобильных приложений, обусловленных ограниченностью энергетических ресурсов мобильных устройств и возрастанием вычислительных требований современных сервисов. Рассмотрены основные источники энергопотребления, такие как локальные вычислительные операции, передача данных по беспроводным каналам, применение алгоритмов искусственного интеллекта и использование мобильных граничных вычислений (MEC). В работе проведён обзор современных алгоритмов и стратегий оптимизации, включающих методы глубокого обучения с подкреплением (DRL), совместное распределение ресурсов (JCC), методы последовательного выпуклого приближения (SCA) и подходы, основанные на модели игры Штакельберга. Полученные результаты демонстрируют потенциал комплексного применения гибридных алгоритмов для повышения энергоэффективности мобильных приложений и открывают перспективы для дальнейших исследований в данной области. Сведения, отраженные в статье, будут представлять интерес для специалистов в области разработки мобильных приложений, системных инженеров и исследователей, заинтересованных в оптимизации энергопотребления посредством инновационных алгоритмических подходов и стратегий программной оптимизации. Данный материал также будет полезен академическим сотрудникам, аспирантам и экспертам в сфере энергоэффективных вычислительных решений, стремящимся интегрировать передовые методологии в создание устойчивых и экономичных мобильных систем.

ABSTRACT

The article analyzes the problems of energy efficiency in the development of mobile applications due to the limited energy resources of mobile devices and the increasing computing requirements of modern services. The main sources of energy consumption are considered, such as local computing operations, data transmission over wireless channels, the use of artificial intelligence algorithms and the use of mobile edge computing (MEC). The paper provides an overview of modern algorithms and optimization strategies, including deep reinforcement learning (DRL) methods, joint resource allocation (JCC), sequential convex approximation (SCA) methods and approaches based on the Stackelberg game model. The results obtained demonstrate the potential of integrated application of hybrid algorithms to improve the energy efficiency of mobile applications and open up prospects for further research in this area. The information reflected in the article will be of interest to specialists in the field of mobile application development, system engineers and researchers interested in optimizing energy consumption through innovative algorithmic approaches and software optimization strategies. This material will also be useful for academic staff, graduate students, and experts in the field of energy-efficient computing solutions seeking to integrate advanced methodologies into the creation of sustainable and cost-effective mobile systems.

Ключевые слова: энергоэффективность, мобильные приложения, мобильные граничные вычисления (MEC), распределение ресурсов, глубокое обучение с подкреплением (DRL), последовательное выпуклое приближение (SCA), игра Штакельберга, оптимизация энергопотребления.

Keywords: energy efficiency, mobile applications, mobile edge computing (MEC), resource allocation, deep reinforcement learning (DRL), sequential convex approximation (SCA), Stackelberg game, energy consumption optimization.

Введение

В условиях массового проникновения смартфонов в повседневную жизнь, проблема энергоэффективности мобильных приложений приобретает особую актуальность. Ограниченные энергетические ресурсы мобильных устройств требуют разработки новых алгоритмов и стратегий, способных снизить энергозатраты при выполнении вычислительных и коммуникационных операций, тем самым продлевая время автономной работы и улучшая качество пользовательского опыта [4,7].

В последние годы наблюдается активное развитие методов повышения энергоэффективности в разработке мобильных приложений, что обусловлено возрастающей потребностью в рациональном распределении вычислительных ресурсов и снижении энергозатрат в условиях динамично развивающихся мобильных сетей. Литература по данной теме условно делится на несколько групп, каждая из которых фокусируется на отдельных аспектах проблемы.

Первая группа исследований посвящена внедрению искусственного интеллекта в процессы разработки мобильных приложений. Так, работы Юрченко Т. В., Цветков И. О. [2] и Zhang Y., Xu Y., Pan Y. [3] анализируют современные тренды применения методов машинного обучения для оптимизации энергопотребления, что позволяет адаптивно настраивать алгоритмы работы приложений в зависимости от состояния системы. Фундаментальные концепции, изложенные в труде Russell S. Norvig P. [11], обеспечивают теоретическую базу для дальнейшей разработки интеллектуальных систем, способных минимизировать затраты энергии за счёт предиктивного анализа и автоматизированного управления ресурсами.

Вторая группа охватывает исследования, направленные на распределение и оптимизацию вычислительных ресурсов в среде мобильных граничных и многоадресных вычислений. В работе Чипсановой Е. В., Елагина В. С. [1] рассматриваются методы распределения ресурсов в концепции мобильных граничных вычислений, что позволяет обеспечить баланс между вычислительной нагрузкой и энергозатратами. Аналогичные проблемы изучаются в исследованиях Bebortta S., Singh A. K., Senapati D. [4] и Liu H. et al. [8], где предлагаются модели совместной оптимизации распределения запросов и вычислительных ресурсов для улучшения энергоэффективности. Важное место занимают исследования Li C., Zhang Y., Luo Y. [5] и Guo S. et al. [12], где для решения задач распределения ресурсов применяются методы глубокого обучения и комбинированные подходы, включающие элементы игровой теории, что позволяет реализовать более гибкое и динамичное управление энергопотреблением. Дополнительно, обзор тенденций и вызовов в мобильных граничных вычислениях, представленный Khan M. A., Ahmadon M. A. [7], демонстрирует, как комплексное рассмотрение инфраструктурных и технологических аспектов способствует поиску оптимальных решений в сфере энергоменеджмента.

Третья группа исследований концентрируется на использовании игровых моделей для решения задач распределения ресурсов и вычислительного оффлоада. Здесь значимый вклад вносит работа Stein A. et al. [6], в которой представлена модель эволюционной Stackelberg-игры, позволяющая учитывать адаптивную природу распределения ресурсов в динамично меняющихся системах. Похожим образом, Hu H. C., Wang P. C. [10] предлагают модель игры offline вычислений для многоканальных беспроводных сенсорных сетей, где анализируются конкурентные и кооперативные стратегии участников, влияющие на энергоэффективность сети. Применение игровых механизмов в данных исследованиях подчёркивает важность стратегического взаимодействия между

элементами системы для достижения оптимальных показателей энергосбережения.

Четвёртая группа литературы представляет альтернативные алгоритмические подходы к решению задач оптимизации, не всегда напрямую связанных с мобильными приложениями, но имеющих потенциал для адаптации к данной области. В работе Chen Y. et al. [9] предложена параллельная не-выпуклая аппроксимационная модель, изначально разработанная для портфельного дизайна с учетом риска, которая может быть трансформирована для решения сложных задач оптимизации распределения ресурсов в мобильных вычислительных системах.

Таким образом, обзор современной литературы показывает, что подходы к обеспечению энергоэффективности в разработке мобильных приложений варьируются от использования интеллектуальных алгоритмов и методов глубокого обучения до сложных моделей распределения ресурсов, основанных на игровых теориях и оптимизационных алгоритмах. Противоречия в литературе проявляются, прежде всего, в различиях между теоретическими моделями и их практической реализуемостью: одни исследования демонстрируют высокую эффективность предлагаемых методов в смоделированных условиях, в то время как другие указывают на сложности интеграции данных решений в реальные системы с динамическими условиями работы.

Цель настоящего исследования состоит в анализе алгоритмов и стратегий, направленных на минимизацию энергопотребления мобильных приложений при сохранении высокого качества обслуживания пользователей.

Научная новизна заключается в предложении нового подхода в процессе разработки мобильных приложений, с учетом энергоэффективности, что стало возможным благодаря анализу других научных публикаций.

Авторская гипотеза заключается в том, что интеграция алгоритмов машинного обучения для управления распределением ресурсов позволит снизить энергопотребление мобильных приложений без компромисса по времени отклика и общей производительности системы.

Методология исследования основывается на сравнительном анализе других научных публикаций.

1. Анализ источников энергопотребления в мобильных приложениях

Эффективное использование энергии является важным аспектом в разработке современных мобильных приложений, поскольку мобильные устройства обладают ограниченными энергетическими ресурсами. Энергозатраты обусловлены рядом взаимосвязанных факторов, включающих локальные вычислительные операции, передачу данных по беспроводным сетям, применение алгоритмов искусственного интеллекта (ИИ), а также использование технологий мобильных граничных вычислений (MEC). Каждый из этих компонентов вносит вклад в общее энергопотребление устройства, что требует комплексного анализа для дальнейшей оптимизации.

Локальные вычислительные операции представляют собой значительный источник энергозатрат. При выполнении сложных вычислительных задач на мобильном устройстве, таких как обработка изображений или анализ больших массивов данных, каждый такт центрального процессора (ЦП) потребляет энергию, что суммарно может приводить к большому расходу [3]. Такие вычислительные операции особенно характерны для приложений, реализующих функции распознавания образов или обработки видео, где высокие требования к ресурсам усугубляют проблему энергопотребления.

Передача данных по беспроводным сетям также влияет на энергозатраты мобильных устройств. Энергопотребление при передаче данных определяется не только объемом передаваемой информации, но и характеристиками сети, такими как скорость передачи, качество сигнала и задержки в канале. Исследования показывают, что нестабильные условия передачи данных и увеличение объема информации ведут к значительному росту энергозатрат [10, 11]. Этот аспект особенно актуален для приложений, которые требуют постоянного обмена данными в режиме реального времени.

Алгоритмы искусственного интеллекта (ИИ), применяемые для автоматизации процессов, анализа и персонализации пользовательского опыта, могут повышать эффективность работы приложений. Однако, выполнение сложных моделей глубокого обучения требует значительных вычислительных ресурсов, что приводит к дополнительному энергопотреблению [11]. Таким образом, хотя ИИ обеспечивает улучшение функциональности, его использование требует оптимизации с целью минимизации энергетических издержек.

Мобильные граничные вычисления (MEC) представляют собой перспективное направление, позволяющее разгрузить мобильное устройство за счет передачи вычислительных задач на внешние серверы, расположенные в непосредственной близости. Этот подход позволяет снизить нагрузку на локальные ресурсы, однако сопряжён с дополнительными энергозатратами на передачу данных по беспроводным каналам [7]. Таким образом, использование MEC создает компромисс между снижением локального энергопотребления и увеличением затрат на коммуникацию.

Для наглядного представления основных источников энергопотребления в мобильных приложениях приведена таблица 1.

Таблица 1

Основные источники энергопотребления в мобильных приложениях [3, 4, 7, 10, 12].

Table 1

Main sources of energy consumption in mobile applications [3, 4, 7, 10, 12].

Фактор	Описание
Локальное вычисление	Выполнение вычислительных операций непосредственно на мобильном устройстве; энергозатраты ЦП при сложных задачах
Передача данных	Энергозатраты, связанные с передачей данных по беспроводным каналам, зависят от объёма и качества сети
Алгоритмы ИИ	Использование моделей глубокого обучения и ИИ для обработки данных, что требует высоких вычислительных ресурсов
Обработка данных в MEC	Передача данных на серверы мобильных граничных вычислений для обработки, что снижает локальную нагрузку, но требует энергии на связь
Системные задержки	Задержки в передаче и обработке данных, ведущие к повторным попыткам передачи и дополнительным энергозатратам

Таким образом можно сделать вывод, что оптимизация энергопотребления мобильных приложений должна учитывать комплексный характер проблемы. Баланс между локальными вычислениями, передачей данных и использованием внешних вычислительных ресурсов является главным фактором для снижения общего энергопотребления. Применение адаптивных алгоритмов, включая методы ИИ, позволяет динамически распределять ресурсы и минимизировать затраты энергии, однако требует тщательной настройки параметров для избежания негативного влияния на производительность устройства.

2. Современные алгоритмы и стратегии оптимизации энергопотребления

В условиях постоянного роста требований к мобильным устройствам и усложнения функционала приложений, современные алгоритмы оптимизации энергопотребления играют важную роль в повышении эффективности работы мобильных систем. Для достижения оптимального баланса между вычислительной производительностью и энергозатратами разработаны комплексные подходы, основанные на методах искусственного интеллекта, адаптивном распределении ресурсов и принципах теории игр. Рассмотрим основные направления современных исследований.

Одним из перспективных подходов является применение алгоритмов глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL) для быстрого управления распределением вычислительных и коммуникационных ресурсов в системах мобильных граничных вычислений (MEC). DRL позволяет моделировать сложные сценарии, где оптимальное решение зависит от меняющихся условий сети, объема данных и вычислительной нагрузки. Например, метод совместного управления коммуникационными и вычислительными ресурсами на основе DRL, реализованный в среде SDN, продемонстрировал снижение времени обслуживания и уменьшение энергопотребления по сравнению с простыми методами [5]. При этом преимуществами являются адаптивность и способность к самообучению, что особенно важно для динамических мобильных условий. Однако высокая вычислительная сложность обучения модели и необходимость большого объёма данных для тренировки остаются серьёзными ограничениями данного подхода [11].

Подход, основанный на совместном распределении коммуникационных и вычислительных ресурсов (Joint Communication and Computation, JCC), направлен на оптимизацию выполнения задач посредством эффективного разделения вычислительной нагрузки между мобильными устройствами и серверными ресурсами MEC. Данный метод позволяет минимизировать энергозатраты за счёт выбора оптимального места для обработки данных – локально или на удалённом сервере, в зависимости от текущей загрузки и характеристик сети. Использование алгоритмов JCC обеспечивает снижение энергопотребления и задержек по сравнению с эвристическими методами, такими как случайный выбор или «грубое жадное» распределение ресурсов [9, 12]. Основным преимуществом является возможность адаптивного распределения ресурсов с учётом ограничений по времени и доступных вычислительных мощностей, однако данная стратегия требует точного моделирования параметров сети и своевременной обратной связи.

Метод последовательного выпуклого приближения (Successive Convex Approximation, SCA) применяется для решения невыпуклых оптимизационных задач, связанных с минимизацией суммарного энергопотребления мобильных устройств при разгрузке вычислительных задач в облачные системы. SCA позволяет эффективно управлять компромиссами между временем передачи данных и затратами энергии, особенно в приложениях дополненной реальности (AR), где задержки критичны для пользовательского опыта. Исследования показывают, что применение SCA в системах MEC обеспечивает до 37% экономии энергии по сравнению с раздельной разгрузкой, за счёт совместного использования каналов связи и вычислительных ресурсов [1, 2]. При этом основным ограничением является необходимость точной оценки параметров канала и

вычислительных нагрузок для корректного применения выпуклого приближения.

Использование принципов теории игр, в частности, модели игры Штакельберга, позволяет выстроить эффективную стратегию распределения вычислительных ресурсов между мобильными устройствами, точками доступа и серверами MEC. В данной модели точка доступа выступает в роли лидера, устанавливая цену и определяя оптимальное количество ресурсов, в то время как мобильные устройства, действуя как последователь, принимают решение о покупке ресурсов с учетом установленных условий. Результаты исследований показывают, что стратегия, основанная на игре Штакельберга, позволяет обеспечить максимальную полезность сервера при одновременном снижении энергопотребления и обеспечении требуемых задержек для конечного пользователя [6]. Основными преимуществами данного подхода являются гибкость в управлении ресурсами и возможность адаптации к изменяющимся условиям сети, однако сложность моделирования взаимодействия агентов требует использования продвинутых аналитических методов.

Ниже в таблице 2 проведен сравнительный анализ современных алгоритмов оптимизации энергопотребления.

Таблица 2

Сравнительный анализ современных алгоритмов оптимизации энергопотребления [3, 5, 7].

Table 2

Comparative analysis of modern algorithms for optimizing energy consumption [3, 5, 7].

Стратегия	Ключевые технологии	Тенденции развития
DRL (Deep Reinforcement Learning)	Глубокое обучение, алгоритмы подкрепления	Оптимизация процесса обучения с целью снижения вычислительной сложности, интеграция с transfer learning и гибридными моделями для повышения эффективности
JCC (Joint Communication and Computation)	Совместное распределение вычислительных и коммуникационных ресурсов	Разработка более динамичных и адаптивных алгоритмов, интеграция с методами DRL для повышения точности оценки сетевых параметров и улучшения механизмов распределения ресурсов
SCA (Successive Convex Approximation)	Последовательное выпуклое приближение для невыпуклых задач	Комбинирование классических методов выпуклого приближения с алгоритмами машинного обучения для повышения точности оценки параметров и ускорения сходимости оптимизационных процессов
Теория игр (Модель Штакельберга)	Моделирование стратегического взаимодействия между агентами	Разработка более сложных адаптивных моделей с использованием алгоритмов машинного обучения для прогнозирования поведения агентов, а также интеграция с новыми технологиями распределения ресурсов

Таким образом, современные алгоритмы оптимизации энергопотребления в мобильных приложениях демонстрируют высокий потенциал за счёт интеграции методов ИИ, адаптивного распределения ресурсов и теоретико-игровых моделей. Комплексный подход, основанный на синергии данных методов, позволяет снизить энергозатраты без ущерба для производительности и качества обслуживания, что является важным в условиях постоянного увеличения функциональной нагрузки на мобильные устройства. Дополнительное использование экспериментальных данных и симуляционных моделей подтверждает эффективность предложенных стратегий и открывает перспективы для дальнейших исследований в данной области.

3. Практическая реализация и перспективы внедрения

Интеграция алгоритмов оптимизации энергопотребления в мобильные приложения становится одной из важных задач в условиях растущих вычислительных требований и ограниченных энергетических ресурсов мобильных устройств. Практическая реализация данных алгоритмов предполагает их адаптацию под конкретные сценарии использования, что требует тесного взаимодействия между разработчиками программного обеспечения, инженерами по сетям и специалистами в области искусственного интеллекта.

Применение методов глубокого обучения с подкреплением (DRL) в системах мобильных граничных вычислений (MEC) позволяет снизить время обслуживания и энергопотребление за счёт быстрого распределения вычислительных и коммуникационных ресурсов [2, 8]. Эксперименты, проведённые в лабораторных условиях и пилотных проектах, подтверждают, что DRL-алгоритмы способны адаптироваться к изменяющимся условиям сети, оптимизируя распределение ресурсов и, как следствие, уменьшая затраты энергии [1,2]. Метод совместного распределения коммуникационных и вычислительных ресурсов (JCC) уже на практике применяется для реализации распределённых систем обработки данных, где задачи делятся между локальными устройствами и удалёнными серверами MEC [1]. Данный подход позволяет оптимизировать не только

энергопотребление, но и задержки при выполнении задач, что особенно актуально для приложений, требующих работы в реальном времени, таких как системы распознавания образов и дополненной реальности.

Метод последовательного выпуклого приближения (SCA) нашёл применение в сценариях, связанных с обработкой больших объёмов данных, например, в AR-приложениях. Эффективность SCA подтверждена экспериментальными данными, где совместное использование каналов связи и вычислительных ресурсов обеспечило до 37% экономии энергии по сравнению с традиционными методами разгрузки [4].

Стратегии, основанные на теории игр, в частности модели игры Штакельберга, также продемонстрировали свою практическую применимость в условиях распределённых вычислительных систем. Использование этой модели позволяет точнее регулировать ценовые механизмы и распределение ресурсов между точками доступа и мобильными устройствами, что приводит к максимизации общей полезности системы и снижению энергозатрат [6].

Практическая реализация данных подходов сопровождается пилотными проектами в различных секторах: от систем «умного дома» до промышленных IoT-сетей. Например, интеграция алгоритмов DRL в системы управления энергопотреблением в «умном доме» позволила снизить общие энергозатраты на 20–30%, а применение модели игры Штакельберга дало возможность оптимизировать распределение вычислительных ресурсов между различными точками доступа, что привело к улучшению качества обслуживания конечных пользователей [6, 12]. Кроме того, современные симуляционные модели, разработанные на базе экспериментальных данных, подтверждают эффективность комплексного применения описанных методов.

Применение гибридных подходов позволяет учитывать динамичность условий эксплуатации мобильных устройств и сетей, обеспечивая адаптивное управление энергопотреблением в реальном времени. Будущие исследования будут направлены на интеграцию алгоритмов машинного обучения с методами теории игр для создания самооптимизирующихся систем распределения ресурсов, способных работать в условиях быстро меняющихся сетевых параметров [4,7].

Ниже представлена таблица 3, обобщающая показатели практической реализации различных алгоритмов оптимизации энергопотребления:

Таблица 3

Ключевые показатели практической реализации алгоритмов оптимизации энергопотребления [1-3].

Table 3

Key indicators of practical implementation of algorithms for optimizing energy consumption [1-3].

Алгоритм/Стратегия	Область применения	Преимущества	Проблемы внедрения
DRL	MEC, адаптивное управление ресурсами в динамичных сетях	Высокая адаптивность, снижение времени обслуживания, динамическая оптимизация энергопотребления	Высокая вычислительная сложность, требовательность к объёму данных для обучения
JCC	Распределённые вычислительные системы, системы реального времени	Эффективное распределение вычислительной нагрузки, снижение задержек и энергозатрат, оптимизация распределения между локальными устройствами и MEC	Необходимость точного моделирования параметров сети, зависимость от обратной связи от системы
SCA	AR-приложения, системы совместной обработки данных	Существенное снижение энергопотребления за счёт совместного использования каналов связи и вычислительных ресурсов	Чувствительность к изменению условий сети, необходимость точной оценки параметров канала
Игра Штакельберга	Системы «умного дома», распределение ресурсов в сетях с множеством участников	Гибкое регулирование ценовых механизмов, оптимизация распределения вычислительных ресурсов, максимизация общей полезности системы при снижении энергозатрат	Сложность моделирования взаимодействия между агентами, необходимость продвинутых аналитических инструментов для прогнозирования поведения

Таким образом, практическая реализация описанных методов демонстрирует их высокую эффективность в реальных условиях эксплуатации. Комплексное применение алгоритмов оптимизации

энергопотребления открывает новые перспективы для разработки мобильных приложений с повышенной энергоэффективностью, что в свою очередь способствует продлению времени автономной работы устройств и улучшению качества обслуживания конечных пользователей. Дальнейшие исследования в этой области будут направлены на разработку гибридных систем, способных адаптироваться к постоянно меняющимся условиям эксплуатации в реальном времени, что является важным фактором для успешного внедрения инновационных технологий в мобильную разработку.

Заключение

В статье проведён детальный анализ источников энергопотребления мобильных приложений и рассмотрены современные алгоритмы и стратегии оптимизации, направленные на снижение затрат энергии без ущерба для производительности систем. Комплексное использование методов глубокого обучения с подкреплением, совместного распределения вычислительных и коммуникационных ресурсов, последовательного выпуклого приближения, а также стратегий, основанных на игре Штакельберга, позволяет создать адаптивные и эффективные системы управления энергозатратами. Практическая реализация данных подходов подтверждена эмпирическими исследованиями и кейс-стади, что свидетельствует о высоком потенциале интеграции оптимизационных алгоритмов в реальные мобильные системы. Полученные результаты не только способствуют продлению времени автономной работы мобильных устройств, но и открывают новые возможности для дальнейших исследований в области энергоэффективной мобильной разработки, в том числе за счёт создания гибридных систем, объединяющих преимущества различных методик. Дальнейшие исследования, направленные на оптимизацию параметров алгоритмов и их адаптацию к динамичным условиям эксплуатации, являются перспективным направлением, способствующим повышению качества и надёжности мобильных приложений в условиях постоянно растущих требований к функциональности и энергоэффективности.

Литература

1.Чипсанова Е. В., Елагин В. С. Методы распределения ресурсов концепции мобильных граничных вычислений //Наукоемкие технологии в космических исследованиях Земли. – 2024. – Т. 16. – №. 1. – С. 4-13.

2.Юрченко Т. В., Цветков И. О. Искусственный интеллект в сфере разработки мобильных приложений //ННГАСУ. – 2024. - С. 160 – 163.

3.Zhang, Y. Artificial Intelligence in Mobile App Development: Current Trends and Future Directions / Zhang, Y., Xu, Y., Pan, Y. // Journal of Software Engineering and Applications. – 2022. – 20 с.

4.Bebortta S., Singh A. K., Senapati D. Performance analysis of multi-access edge computing networks for heterogeneous IoT systems //Materials Today: Proceedings. – 2022. – Т. 58. – С. 267-272.

5.Li C., Zhang Y., Luo Y. Deep reinforcement learning-based resource allocation and seamless handover in multi-access edge computing based on SDN //Knowledge and Information Systems. – 2021. – Т. 63. – №. 9. – С. 2479-2511.

6.Stein A. et al. Stackelberg evolutionary game theory: how to manage evolving systems //Philosophical Transactions of the Royal Society B. – 2023. – Т. 378. – №. 1876. – С. 2-10.

7.Khan M. A., Ahmadon M. A. Trends and Challenges in Mobile Edge Computing for the Next Generation Massive Internet of Things. – 2023. - Vol. 4. - pp. 28-42.

8.Liu H. et al. Joint optimization of request assignment and computing resource allocation in multi-access edge computing //IEEE Transactions on Services Computing. – 2022. – Т. 16. – №. 2. – С. 1254-1267.

9.Chen Y. et al. A parallel non-convex approximation framework for risk parity portfolio design //Parallel Computing. – 2023. – Т. 116. – С. 1-9.

10.Hu H. C., Wang P. C. Computation offloading game for multi-channel wireless sensor networks //Sensors. – 2022. – Т. 22. – №. 22. – С. -2-8.

11.Russell, S. Norvig, P. Artificial Intelligence: A Modern Approach / S. Russell, P. Norvig. – London : Pearson, 2020. –С.20-34.

12.Guo S. et al. Mobile edge computing resource allocation: A joint Stackelberg game and matching strategy //International Journal of Distributed Sensor Networks. – 2019. – Т. 15. – №. 7. – С. 1-15.

References

1.Chipsanova E. V., Elagin V. S. Resource allocation methods for mobile edge computing concepts //High-tech technologies in space exploration of the Earth. – 2024. – Vol. 16 (1). – pp. 4-13.

2.Yurchenko T. V., Tsvetkov I. O. Artificial intelligence in the field of mobile application development //NGASU. – 2024. - pp. 160-163.

3.Zhang, Y. Artificial Intelligence in Mobile App Development: Current Trends and Future Directions / Zhang, Y., Xu, Y., Pan, Y. // Journal of Software Engineering and Applications. – 2022. – pp. 20 .

4.Bebortta S., Singh A. K., Senapati D. Performance analysis of multi-access edge computing networks for heterogeneous IoT systems //Materials Today: Proceedings. – 2022. – Vol. 58. – pp. 267-272.

5.Li C., Zhang Y., Luo Y. Deep reinforcement learning-based resource allocation and seamless handover in multi-access edge computing based on SDN //Knowledge and Information Systems. – 2021. – Vol. 63 (9). – pp. 2479-2511.

6.Stein A. et al. Stackelberg evolutionary game theory: how to manage evolving systems //Philosophical Transactions of the Royal Society B. – 2023. – Vol. 378 (1876). – pp. 2-10.

7.Khan M. A., Ahmadon M. A. Trends and Challenges in Mobile Edge Computing for the Next

Generation Massive Internet of Things. – 2023. - Vol. 4. - pp. 28-42.

8.Liu H. et al. Joint optimization of request assignment and computing resource allocation in multi-access edge computing //IEEE Transactions on Services Computing. – 2022. – Vol. 16 (2). – pp. 1254-1267.

9.Chen Y. et al. A parallel non-convex approximation framework for risk parity portfolio design //Parallel Computing. – 2023. – Vol. 116. – pp. 1-9.

10.Hu H. C., Wang P. C. Computation offloading game for multi-channel wireless sensor networks //Sensors. – 2022. – Vol. 22. – pp. -2-8.

11.Russell, S. Norvig, P. Artificial Intelligence: A Modern Approach / S. Russell, P. Norvig. – London : Pearson, 2020. – pp. 20-34.

12.Guo S. et al. Mobile edge computing resource allocation: A joint Stackelberg game and matching strategy //International Journal of Distributed Sensor Networks. – 2019. –Vol. 15 (7). – pp. 1-15.

ТЕХНИЧЕСКИЕ НАУКИ

COMPARISON OF ADAPTIVE LEAST MEAN SQUARE FILERS FOR RADAR SIGNAL PROCESSING

Trung Thanh Nguyen
Ph-D, Department of Electronic Warfare,
Le Quy Don Technical University,
Vietnam

SUMMARY

In radar systems, accurate estimation of signal parameters plays a vital role in correctly identifying radar sources and providing a high level of electronic warfare map. On the other hand, the signal at the receiver consists of echo and noisy signals. Therefore, to improve accuracy when estimating signal parameters, this article studies the adaptive least mean square algorithms (LMS) for reducing noise in digital signal processing. The effectiveness of the LMS filters is evaluated through simulation of all variants of LMS, such as normalized LMS (NLMS) and complex LMS (CLMS), sign LMS (SLMS), with simulated radar signals such as radar pulses, linear frequency modulation, and Barker code in a MATLAB environment. Simulation results show that the LMS filter has better noise filtering than the LMS and its variants, such as the NLMS and LLMS.

Keywords: adaptive filter, least mean square algorithm, radar signal, MATLAB environment.

1. Introduction

In modern warfare, passive radar systems play an important role, and the signal processing results of passive systems provide a comprehensive picture of the electromagnetic field and support the commander in making timely decisions to improve combat effectiveness and victory ability. The main tasks of passive radar systems include:

- Orientation and location of emission sources,
- Classification and identification of emission sources [1, 2].

On the other hand, to solve the problem of classifying and identifying radar sources, the requirements for passive radar systems are as follows: measurement and analysis of signals with high accuracy. Standard methods used for signal processing include three main groups: The first group is signal processing in the time domain, the second group of methods is in the frequency domain, and the last group of solutions is in both the time and frequency domains [3, 4].

Typical frequency domain signal processing methods: fast Fourier transform (FFT) or Z-transform [5, 6]. The limitation of this group of methods is that it can only observe signals in the frequency domain, and it is difficult to determine the variation of signal frequency over time or, in other words, to identify modulated signals. To overcome the above limitations, signal processing methods in both time and frequency domains are used. These methods provide instantaneous information about the frequency change over time. The results of this analysis are called time-frequency images of the signal and are used as input to the recognition units. Common signal processing methods in both time-frequency domains include short-time Fourier transform (STFT) [7], Wigner-Ville distribution (WVD) [8], or continuous Wavelet transform (CWT) [9]. The limitations of these methods are that they require a large processing time and are difficult to process for low-power signals.

On the other hand, to increase the accuracy of the estimation, passive systems are often used in combination with time-domain processing methods such as low-pass filters, high-pass filters, or band-pass filters [10] to remove unwanted frequencies. Furthermore, passive systems also use digital filters such as FIR and IIR to limit the impact of noise and increase the signal-to-noise ratio (SNR) [11]. Currently, in addition to FIR and IIR digital filters, the system also uses other filters such as Wiener [12], LMS [13], and variations of LMS such as Normalized LMS (N-LMS), or sign LMS (S-LMS), sign-sign LMS (SS-LMS) and signed-regressor LMS (SR-LMS). On the other hand, no studies have compared the performance of LMS algorithms. So, this article compares the performance of the above-mentioned LMS algorithms by MATLAB environment with different types of radar signals such as radar pulse, linear frequency modulation (LFM), and binary phase shift keying with Barker code (Barker code).

Section 2 presents the theoretical descriptions of LMS algorithms. Section 3 shows the simulation results, and Section 4 summarizes the main conclusions.

2. Theoretical description of the least mean square algorithm

An adaptive filter is a linear filter system with a transfer function controlled by variable parameters and a means to adjust those parameters according to an optimization algorithm. Because of the complexity of the optimization algorithms, almost all adaptive filters are digital filters. The block diagram of adaptive filters is shown in Fig. 1.

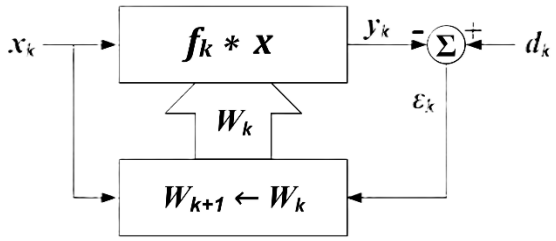


Fig. 1. Block diagram of adaptive filters.

The idea behind a closed-loop adaptive filter is that a variable filter is adjusted until the error is minimized. The Least Mean Squares (LMS) filter and the Recursive Least Squares (RLS) filter are adaptive filters.

2.1Least-mean-square algorithm

The LMS was developed by Widrow and Hoff in 1960. LMS's main idea is to mimic a desired filter by finding the filter coefficients that produce the least mean square of the error signal (the difference between the desired and the actual signal). The parameter of LMS algorithm includes: $x(n), d(n)$ are input and desired signals, M is the number of filter coefficients and μ is the step size factor. The LMS algorithm has the following most important properties:

1. Its form is simple as well as its implementation and capable of delivering high performance during the adaptation process
2. It includes a step-size parameter, which must be selected properly to control the stability and convergence speed of the algorithm.
3. It is stable and robust for a variety of signal conditions.

The block diagram of the LMS algorithm is shown, and its step is written in

Algorithm 1.

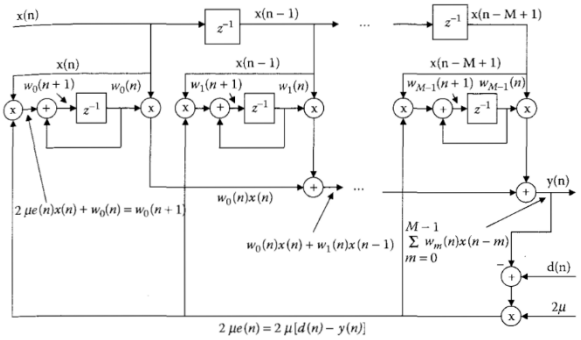


Fig. 2. Block diagram of the LMS algorithm

2.2 Sign least-mean-square algorithm

The first modification of LMS is called the sign algorithm, which is defined by (1), and the algorithm is written in Algorithm 2.

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu \times sign\big(e(n)\big)\mathbf{x}(n) \tag{1}$$

where $sign(x)$ is the signum function. It is defined by (2):

$$sign(n) = \begin{cases} 1 & n > 0 \\ 0 & n = 0 \\ -1 & n < 0 \end{cases} \tag{2}$$

Algorithm 1. LMS algorithm

Parameters:	M = number of filter coefficients
	μ = step – size factor
	$\boldsymbol{x}(n) = [x(n)\, x(n-1)\, \cdots\, x(n-M+1)^T]$
Initialization:	$\boldsymbol{w} = 0$
Computation:	For $n = 0,1,2,\cdots$
	1.$y(n) = \boldsymbol{w}^T(n)\boldsymbol{x}(n)$
	2.$e(n) = d(n) - y(n)$
	3.$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + 2\mu \times e(n)\boldsymbol{x}(n)$

Algorithm 2. Sign LMS algorithm

Parameters:	M = number of filter coefficients

	μ = step – size factor
	$\boldsymbol{x}(n) = [x(n)\ x(n-1)\ \cdots\ x(n-M+1)^T]$
Initialization:	$\boldsymbol{w} = 0$
Computation:	For $n = 0,1,2,\cdots$
	4.$y(n) = \boldsymbol{w}^T(n)\boldsymbol{x}(n)$
	5.$e(n) = d(n) - y(n)$
	6.$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + 2\mu \times sign\big(e(n)\big)\boldsymbol{x}(n)$

2.3 Signed-regressor Least-mean-square algorithm

The second modification of LMS is the signed regressor algorithm, which is written by (3) and Algorithm 3.

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu \times sign\big(\mathbf{x}(n)\big)e(n) \tag{3}$$

where the signum function is applied to $\boldsymbol{x}(n)$.

2.4 Sign-sign least-mean-square algorithm

The last modification of LMS is the sign-sign algorithm, where the signum function is applied to both elements $\boldsymbol{x}(n)$ and $e(n)$. The sign-sign algorithm is defined by (4) and Algorithm 4.

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu \times sign\big(\mathbf{x}(n)\big)sign\big(e(n)\big) \tag{4}$$

Algorithm 3. Signed-regressor LMS algorithm

Parameters:	M = number of filter coefficients
	μ = step – size factor
	$\boldsymbol{x}(n) = [x(n)\ x(n-1)\ \cdots\ x(n-M+1)^T]$
Initialization:	$\boldsymbol{w} = 0$
Computation:	For $n = 0,1,2,\cdots$
	1.$y(n) = \boldsymbol{w}^T(n)\boldsymbol{x}(n)$
	2.$e(n) = d(n) - y(n)$
	3.$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + 2\mu \times sign\big(\boldsymbol{x}(n)\big)e(n)$

Algorithm 4. Sign-sign LMS algorithm

Parameters:	M = number of filter coefficients
	μ = step – size factor
	$\boldsymbol{x}(n) = [x(n)\ x(n-1)\ \cdots\ x(n-M+1)^T]$
Initialization:	$\boldsymbol{w} = 0$
Computation:	For $n = 0,1,2,\cdots$
	1.$y(n) = \boldsymbol{w}^T(n)\boldsymbol{x}(n)$
	2.$e(n) = d(n) - y(n)$
	3.$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + 2\mu \times sign\big(\boldsymbol{x}(n)\big)sign\big(e(n)\big)$

Table 1

Simulation parameters

Parameter	Value
Coefficient of filter M	16
Step size factor μ	0.01
Type of noise	White Gaussian noise
Signal to noise ratio SNR (dB)	0 ÷ 15

Table 2

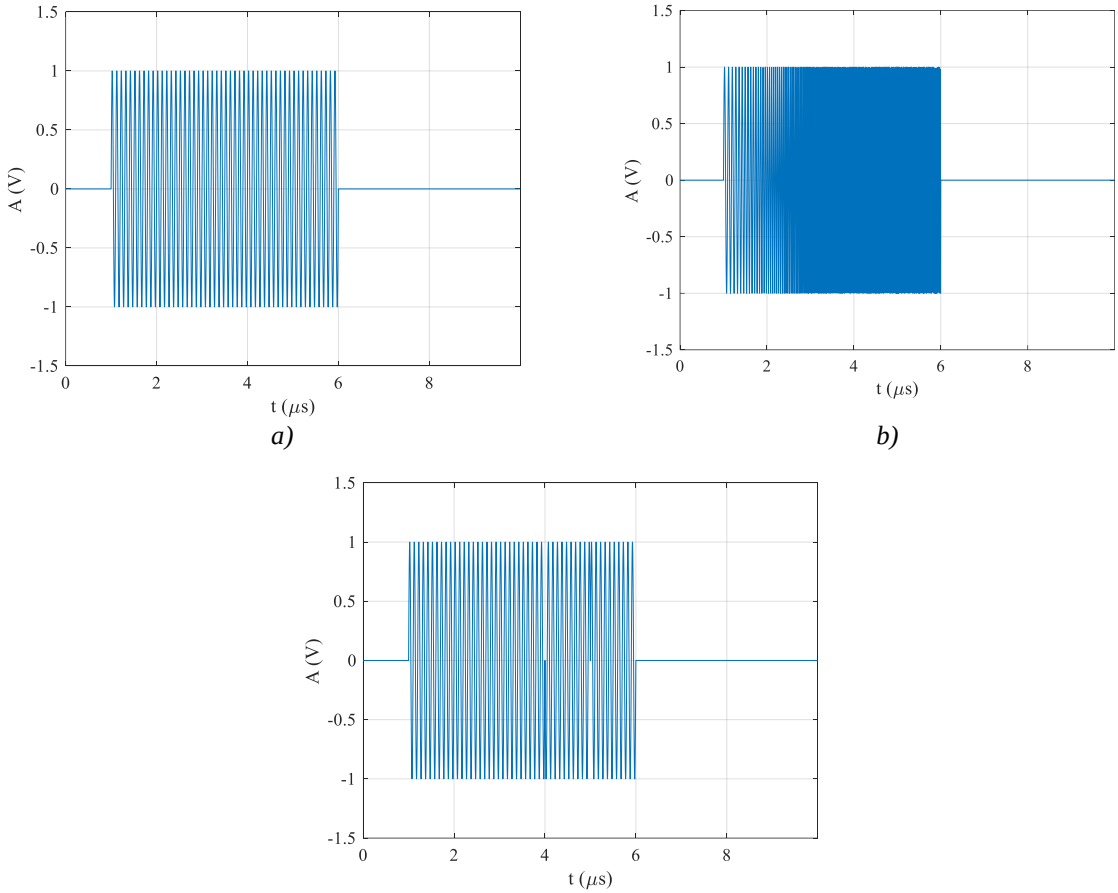
Parameters of simulated radar signals

Signal	Parameters	Value
LFM	Frequency bandwidth BW (MHz)	50
Barker code	Length code N	5
	Sample rate f_s (MHz)	1000
	Pulse width $\tau(\mu s)$	5
	Carrier frequency f_c (MHz)	5

3. Simulation results

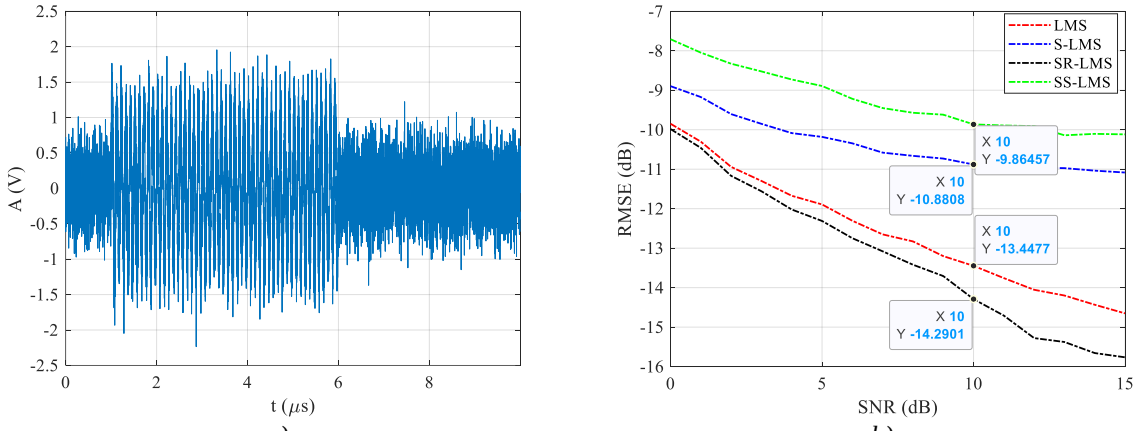
This section analyzes the above-mentioned LMS algorithms in a MATLAB environment with different types of radar signals, such as radar pulse, linear frequency modulation (LFM), and Barker, in the range of

SNR from 0 dBm to 15 dBm. The simulation parameters are shown in Table 1. The parameters of simulated signals are listed in Table 2, and the original signals are shown in Fig. 3.



a) *b)* *c)*

Fig. 3. The original simulated signals: a) radar pulse; b) LFM; c) Barker.



a) *b)*

Fig. 4. Simulation results of radar pulse: a) received signal; b) RMSE versus SNR.

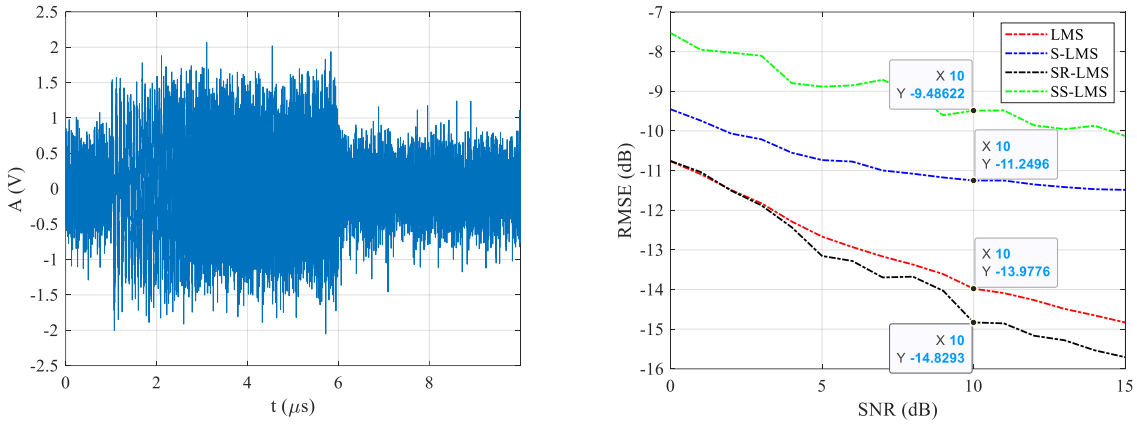
3.1Radar pulse signal

This subsection compares all LMS algorithms' performance in analyzing radar pulses. Fig. 4(a) shows the received signal with SNR = 0 dB. The root means square error (RMSE) of all algorithms versus SNR is shown in Fig. 4(b). It is seen that SR-LMS performs the best results (black line) after LMS (red line) and S-LMS (blue line), and the lowest is provided by SS-LMS (green line).

At SNR = 10 dB, the SR-LMS's RMSE is -14.29 dB, LMS's RMSE is -13.45 dB, and SS-LMS's is -9.87 dB.

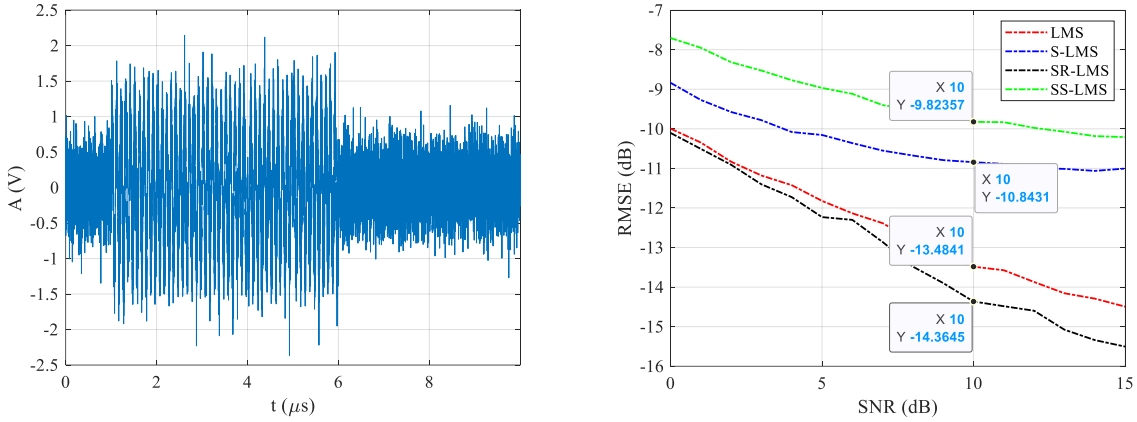
3.2 Linear frequency modulated signal

Fig. 5 shows the performance of all LMS algorithms in analyzing LFM signals with the same step. For example, in analyzing radar pulses, the SR-LMS provides the lowest RMSE (RMSE = -14.83 dB at SNR = 10 dB, black line), and the SS-LMS supplies the highest RMSE (RMSE = -9.49 dB, green line).



a) *b)*

Fig. 5. Simulation results of LFM signal: a) received signal; b) RMSE versus SNR.



a) *b)*

Fig. 6. Simulation results of Barker code: a) received signal; b) RMSE versus SNR.

3.3 Barker coded signal

Fig. 6 shows the performance of all LMS algorithms in analyzing Barker-coded signals. For example, in analyzing radar pulses, the SR-LMS provides the lowest RMSE (RMSE = -14.36 dB at SNR = 10 dB, black line), and the highest RMSE is provided by the SS-LMS (RMSE = -9.83 dB, green line).

All simulation results show that the SR-LMS performs the lowest RMSE in analyzing all radar signals. Also, the SR-LMS provides the best results for analyzing the LFM signal (RMSE = -14.83 dB), BPSK (RMSE = -14.36 dB), and radar pulse (RMSE = -14.29 dB).

4 Conclusion.

This study compared various LMS adaptive filtering algorithms—LMS, Sign LMS (S-LMS), Signed-Regression LMS (SR-LMS), and Sign-Sign LMS (SS-LMS)—for radar signal processing. The SR-LMS algorithm consistently delivered the lowest RMSE across all signal types (radar pulses, LFM, and Barker coded signals) and SNR levels, outperforming other variants. At SNR = 10 dB, SR-LMS achieved RMSEs of -14.29 dB (radar pulse), -14.83 dB (LFM), and -14.36 dB (Barker code). In contrast, SS-LMS showed the poorest performance, with the highest RMSE values, indicating its inefficiency in noise reduction. While LMS and S-LMS performed adequately, SR-LMS proved to be the most effective in dynamic environments, offering superior noise suppression. In conclusion, SR-LMS is the best choice for radar signal processing, particularly in environments with varying signal conditions. Future work will involve testing with real-world signals and hardware implementations.

References

[1]V. Kubecek and P. Svoboda, “Passive Surveillance System for Air Traffic Control”, *1998 28th European Microwave Conference*, Amsterdam, Netherlands, 1998, pp. 546-551, doi: 10.1109/EUMA.1998.338047.

[2]H. D. Griffiths and J. B. Christopher, “An Introduction to Passive Radar”, 2nd edition, Artech House, UK, 2022. ISBN: 9781630818418.

[3]S. K. Mitra, “Digital Signal Processing: A Computer-Based Approach”, 4th edition, McGraw-Hill, 2011, ISBN: 978-0-07-338049-0.

[4]L. N. Trefethen, “Spectral Methods in MATLAB”, SIAM, Philadelphia, USA, 2011. ISBN: 978-0-89871-465-4.

[5]A. Graham, “Communications, Radar and Electronic Warfare”, John Wiley and Sons, West Sussex, UK, 2011. ISBN: 9780470688717.

[6]M. A. Richards, “Fundamentals of Radar Signal Processing”, 3rd edition, McGraw-Hill, 2022, ISBN: 9781260468717.

[7]H. Zhivomirov, “On the Development of STFT-analysis and ISTFT-synthesis Routines and their

Practical Implementation", TEM Journal, vol. 8, no. 1, pp. 56-64, 2019. DOI: 10.18421/TEM81-07.

[8]J. Y. Chen, and B. Z. Li, "The Short-time Wigner-Ville Distribution", Signal Processing, vol. 219, no. 109402, 2024. https://doi.org/10.1016/j.sigpro.2024.109402.

[9]M. Walenczykowska, and A. Kawalec, "Application of Continuous Wavelet Transform and Artificial Naural Network for Automatic Radar Signal Recognition", *Sensors* 22, no. 19: 7434. 2022, https://doi.org/10.3390/s22197434.

[10] J. M. Giron-Sierra, "Digital signal processing with MATLAB examples, volume 1", Springer, New York, USA, 2017. ISBN: 978-981-10-2533-4.

[11] A. Chandra, and S. Chattopadhyay, "Design of Hardware Efficient FIR filter: A Review of the State-of-art Approaches", Engineering Science and Technology, vol. 19, no. 1, pp. 212-226, 2016. https://doi.org/10.1016/j.jestch.2015.06.006.

[12] A. Pathan, and T. D. Memon, "A Correlation-less Approach Towards Adaptive Channel Equalizer Based on Wiener-Hopf Equation", Wireless Personal Communications, vol. 118, no. 4, 2021, pp. 3539-3548. https://doi.org/10.1007/s11277-021-08193-w.

[13] I. Ushenina, "FPGA Implementation of LMS Adaptive Filters Using High-Level Synthesis Tools", In SMART Automatics and Energy. Smart Innovation, Systems and Technologies, vol. 272, Springer, Singapore, 2022. https://doi.org/10.1007/978-981-16-8759-4_47.

[14] S. Zhao, J. Xu, and Y. Zhang, "A Variable Step-size Leaky LMS Algorithm", Wireless Communications and Mobile Computing, vol. 2021, no. 7951643, 11, 2021. https://doi.org/10.1155/2021/7951643.

IMPROVING THE CONTRAST OF TARGET FROM BACKGROUND CLUTTER IN POLARIMETRIC RADAR IMAGE BY USING THE MEAN POLARIMETRY ELLIPTICITY

Pham Trong Hung, Nguyen Tien Thai
Military Technical Academia, PhD, MA,
Vietnam Republic

ABSTRACT

This paper proposes a new method for improving the contrast of target on the background clutter by using the mean of polarimetric ellipticity. Instead of averaging signal samples within a radar cell to produce a polarimetric ellipticity coefficient for detection, this method calculates polarimetric ellipticity coefficients for every signal samples within a radar cell. These coefficients are then averaged to produce the mean polarimetric ellipticity coefficient for detection in the radar cell. Simulation results of the method shows a significant improvement in the contrast of the target on the background clutter in the radar image, and an increase in the probability of target discrimination from clutter background.

1. Introduction

The problem of radar target detection from background clutter using polarimetric parameters has been investigated in many researchs [1], [2], and [3]. In [4] and [5], Kozlov A.I experimentally demonstrates the polarization track effect by the polarimetric ellipticity coefficient K on the circular polarization basic. Later, paper [6] proposes an algorithm of detecting target on the background clutter using polarimetric ellipticity coefficient K. The algorithm, however, produces high false alarm rate and low probability of detection. This is due to the strong fluctuation of the ellipticity coefficient K for the sea clutter, which has the range [-1: +1] and a large deviation [4]. Consequently, the radar images have many speckles of sea clutter, which degrade the image quality and cause difficulties in target discrimination on the background clutter.

There have been several attempts to increase the image quality in the field of polarimetric radar. In [7], Swartz A. A uses the optimal polarimetric filter to impove the contrast of target from background. In this paper a systematic approach is presented for obtaining the optimal polafimetfic matched filter, which produces maximum contrast between two scattering classes. In [8], the authors implement the polarimetric whitening filter (PWF) using Horizontal-Horizontal (HH), Horizontal-Vertical (HV), and Vertical-Vertical (VV) components of Synthetic Aperture Radar (SAR) image, thereby decreasing the speckle in the image.

This paper proposes a new approach to increase the contrast of target on the background clutter. Instead of averaging signal samples within a radar cell to produce a polarimetric ellipticity coefficient (K) for detection, the new method calculates polarimetric ellipticity coefficients for every signal samples within a radar cell. These coefficients are then averaged to produce the mean polarimetric ellipticity coefficient Km for detection in the radar cell. The averaging process reduces the deviation of K parameter, thereby narrowing down its probability density function (PDF). Initial simulation results of the method show a significant improvement in the contrast of the target on the background clutter in the radar image, which leads to an increase in the probability of target discrimination from clutter background.

The aim of this paper is to investigate the improvement in the contrast of polarimetric radar image by using the mean polarimetric ellipticity coefficient. The remainder of this paper is organized as follows: section 2 briefly reviews the radar target detection based on polarimetric parameter K, section 3 proposes a new algorithm for target detection using the mean of ellipticity coefficient Km. Section 4 presents

and compares simulation results of radar target detection using K and Km coefficients. The conclusion is provided in section 5.

2. Review of radar detection based on the polarimetric ellipticity cofficient

The algorithm of detecting target on the background clutter using polarimetric ellipticity coeficient *K* is presented in [6]. In that, the radar system transmits right hand circular polarization (RHCP) signals, and receives both left hand circular polarization (LHCP) and RHCP signals as described in [9], [10].

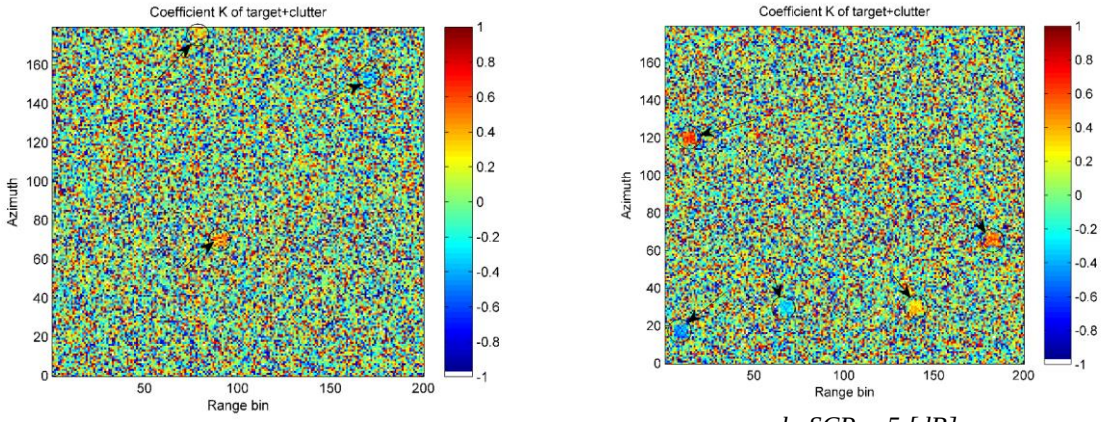
The system measures the circular polarization ratio of a scattered signal as $\left|\dot{P}^{\mathrm{RL}}(t)\right| = \left|\dot{E}^{R}(t)\right| / \left|\dot{E}^{L}(t)\right|$, and then calculates the polarimetric ellipticity coefficient as follows:

$$|K(t)| = \frac{|\dot{P}^{\mathrm{RL}}(t)|-1}{|\dot{P}^{\mathrm{RL}}(t)|+1}; -1 \le K(t) \le 1 \quad (1)$$

The polarimetric ellipticity coefficient *K* from equation (1) is then used for target detection. The simulation results of this algorithm are showed in the Figure 1. In the simulations, 5 targets are randomly generated with ellipticity coefficients are -0.4595, 0.6667, 0.4286, 0.8018 and -0.7143, respectively. Two values of signal-to-clutter ratio (SCR), 0dB and 5 dB, are simulated, which represents two scenarios of normal and strong noise situations. The sea clutter is assumed to be Weibull distribution as described in [11].

Figure 1 shows the measured coefficient *K* of reflected signal from target and clutter in each radar cell. The range of *K* is $K = [-1 \div 1]$ and is color coded, from green (*K*= -1) to red (*K*=1). In the simulation in Figure 1a where SCR = 0 dB, 3/5 targets can be visible. In Figure 1b where SCR= 5 dB, 5/5 target can be visible. Except target point marks, all others points in the radar image represent the ellipticity coefficients of background clutter.

The radar images in the Figure 1 also show the strong fluctuation in the ellipticity coefficient *K* of clutter, which makes it difficult to discriminate the target from background clutter. The distribution of the ellipticity coefficient *K* for clutter background only (without target) is showed in the Figure 2. The distribution is received by calculating the ellipticity coefficients *K* from randomly generated sea-clutter for 20.000 times.



a, SCR = 0 [dB] *b, SCR = 5 [dB]*

Figure 1. The detection of target on the background clutter using polarimetric ellipticity coefficient K

As can be seen in Fig 2, the distribution of the ellipticity coefficient *K* of background clutter spreads in the entire domain of *K* from -1 to +1. This explains the strong fluctuation of *K* parameter for background clutter. If this fluctuation can be reduced, the capability of target discrimination from background clutter can be enhanced as well.

Table 1 shows the experimental results in previous researches [4], [12] with the RCS from -10 dB to 5 dB, depending on the sea clutter conditions. In that, the mean value *K* of sea clutter only is $\langle K \rangle \approx 0$, while that of sea clutter + target is $\langle K \rangle \approx -0.8$. The standard deviation $\sigma_K = 0.23 \div 0.56$ for the sea clutter, and $\sigma_K = 0.07 \div 0.08$ for target + clutter.

It is expected that if the distribution of *K* for background clutter is narrowed down, the radar image quality could be improved. One way of achieving this is to use the mean coefficient *K*, thereby reducing the fluctuation of *K* parameter for background clutter. Detail of the new method of using the mean coefficient *K* for the target detection is introduced in the next section.

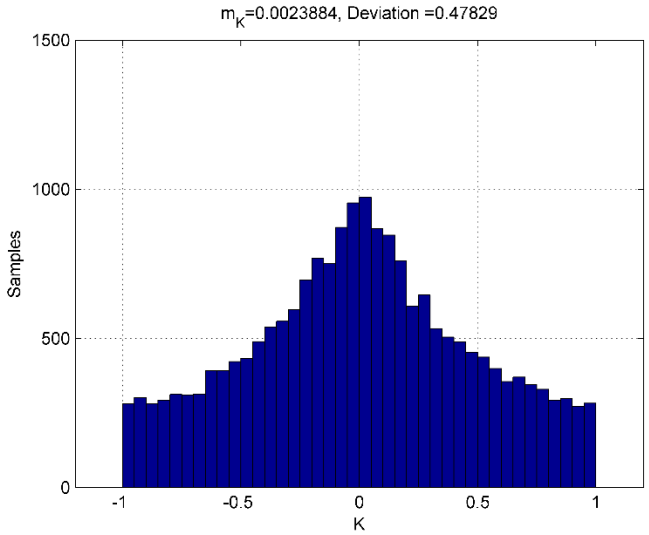


Figure 2. The distribution of the ellipticity coefficient K of clutter background with Weibull model

3. Reducing fluctuation of ellipticity coefficient for background clutter

Assume in each radar cell, *N* samples of radar signal are received $s_1, s_2, \ldots s_N$. In common polarimetric radars, those *N* signal samples are averaged to produce only one value, which represents signal of that radar cell. That value is then used to calculate the ellipticity coefficient *K*, which later can be used for radar detection.

In the proposed algorithm, each of those *N* samples, *S*(*i*), *i* = 1 to *n*, is used to calculate the ellipticity coefficient of its own *K*(*i*), *i* = 1 to *n*. For simplicity, we can rewrite as $K_1, K_2 \ldots K_N$. The mean of those coefficients can be calculated as:

$$K_m = \frac{1}{N}\sum_{i=1}^{N} K(i) \quad (2)$$

This mean value is then used for radar detection.

Simulation results of Weibull model clutter with *N*=10 are presented in Figure 3, 4. In Figure 3, the upper figure emulates the previous algorithm in Section 2, where one value of *K* is calculated for detection in one radar cell. The lower figure emulates the proposed algorithm, where K_m in each radar cell is calculated and used for detection. The distributions of *K* and K_m are illustrated in Figure 4.a and 4.b, respectively.

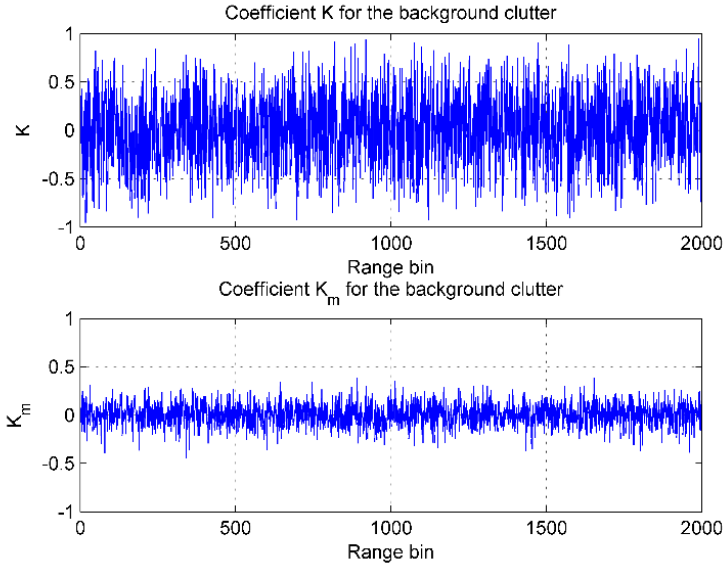


Figure 3. The polarimetric ellipticity coefficient K and K_m for sea clutter

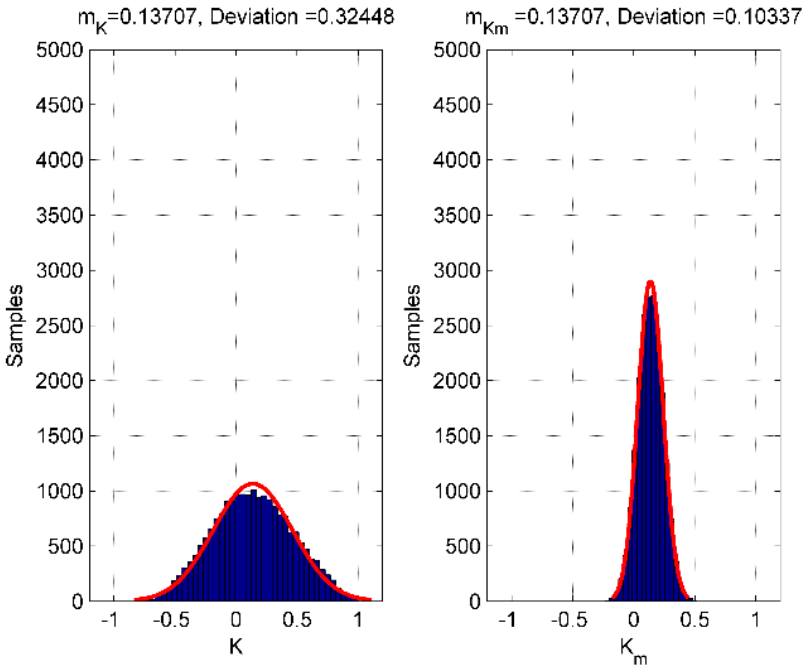


Figure 4. Comparision of the distribution of K and K_m

As can be seen from Figure 3 and 4, the standard deviation of K for background clutter decreases significantly from σ_K=0.32448 to σ_{Km}=0.10337, which is about 5 dB, while their mean values almost unchanged at around zero.

The decrease in the deviation of K for the background clutter means the increase in the target probability detection on the background clutter. When N=10, the deviation of K_m is much smaller comparing to the deviation of K in the Table 1.

Table 1

Experimental results of polarization track of the sea surface [4]

Object	Wave height	Mean value	σ_K
Sea surface	≈ 0.2 m	$\langle K\rangle = -0.2 \div 0.1$	$\sigma_K = 0.23$
Sea surface together with a small object	≈ 0.2 m	$\langle K\rangle = -0.8$	$\sigma_K = 0.07 \div 0.08$
Sea surface	≈0.4÷0.5 m	$\langle K\rangle = 0$	$\sigma_K = 0.26$
Sea surface together with a small object	≈ 0.5 m	$\langle K\rangle = -0.75$	$\sigma_K = 0.033$
Sea surface	≈1.2÷1.5 m	$\langle K\rangle = 0$	$\sigma_K = 0.56$
Sea surface together with a small object	≈1.2÷1.5 m	$\langle K\rangle = -0.7$	$\sigma_K = 0.11 \div 0.125$

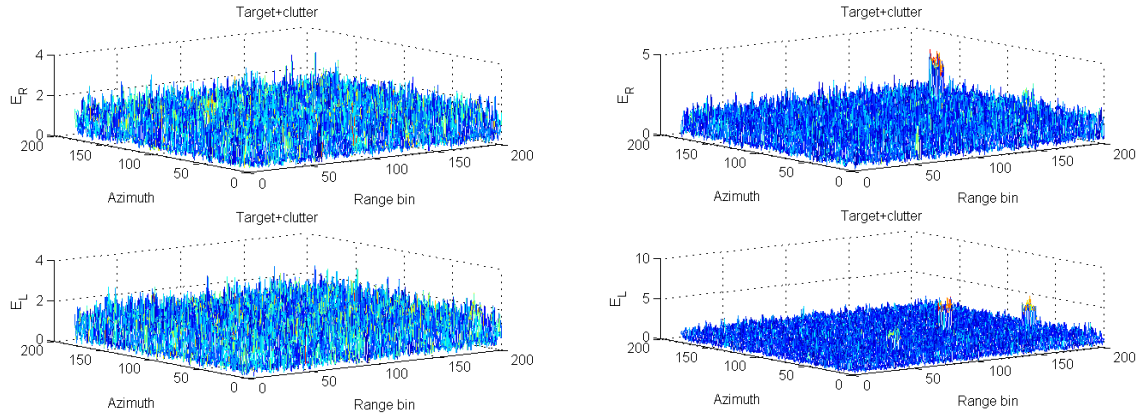
4.Simulation results

Radar image using the polarimetric ellipticity K and K_m coefficients

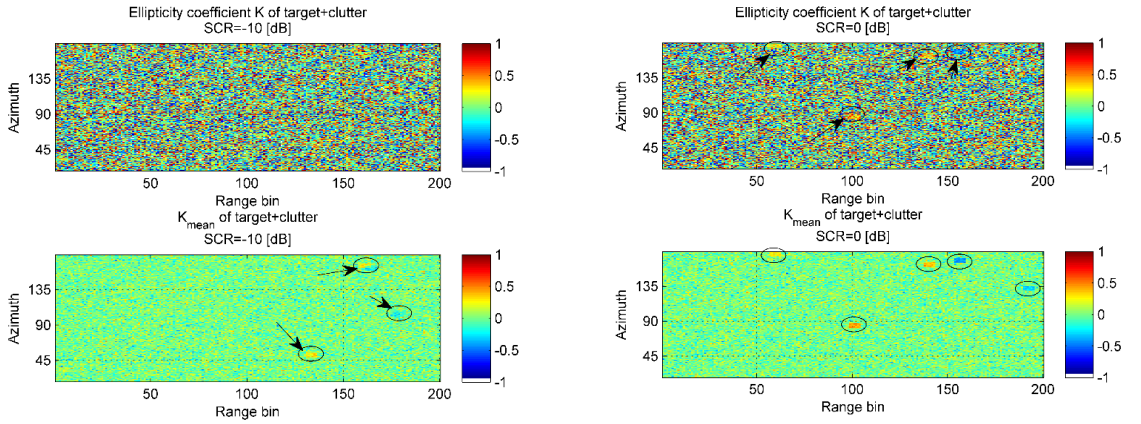
In this section the simulation of target detection using polarimetric ellipticity coefficient is organized in the same manner as in Section 2, except for both K and K_m coefficients. Results in both cases are then compared to prove the improvement of the new algorithm.

In the simulation, 5 targets are generated randomly both in range and azimuth angle, their coefficients K_T are -0.4595; 0.6667; 0.4286; 0.8018; -0.7143 respectively. The background clutter is assumed Weibull model as described in [11]. The SCR is generated with two values -10 and 0 dB to cover its typical range, as described in the experimental researches in [4] and [12]. The results are illustrated in the Figure 5.

In the Figure 5, the upper panels are the signals from LHCP and RHCP of target+clutter; and lower panels are radar image using polarimetric parameter: using K and K_m respectively. As can be seen from Figure 5, if K is used, in Figure 5a where SCR= -10 dB no target can be discriminated from clutter in the radar image. In Figure 5b where SCR= 0 dB, 4/5 targets can be "assumed". On the other hand, if K_m is used, the contrast of target on the background clutter in the radar image is increased significantly, making the targets visible. In particular, in Figure 5a where SCR = -10dB, 3/5 targets can be observed, while at SCR = 0dB as in Figure 5b, all 5 targets can be seen over the clutter background. The more the SCR, the higher contrast of the target over the clutter background in the radar image. With the proposed method, the targets with as low SCR as -10 dB can be visually observed in the radar image.



Signal of target+clutter (SCR=-10 dB) *Signal of target+clutter (SCR=0 dB)*



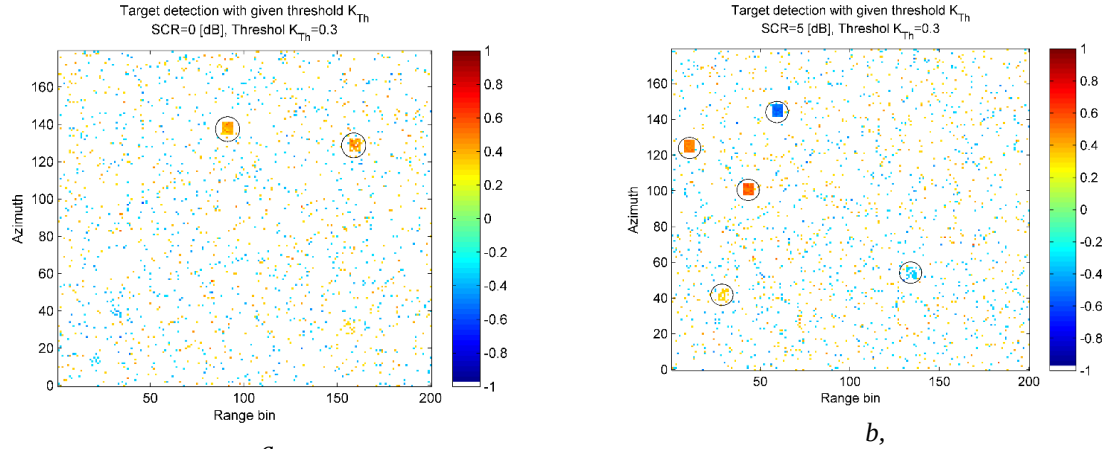
a, *b,*

Figure 5. Simulation results of target detection using polarimetric ellipticity coefficient. Upper panels: signal of target+clutter; and lower panels using K and using K_m

4.2. Further improve radar image by applying threshold K_{Th}

If the detection threshold K_{Th} is applied, a considerable part of the background clutter will be removed from the radar image, making the targets are more visible for the radar operator. This process, however, runs the risk of missing targets if the threshold is set too high. Several simulation results are presented in the Figure 6.

As can be seen from Figure 6 where the threshold is set at K_{Th}= *0.3*, the background clutter and radar signals below that threshold are removed from the image, leaving only those background clutter and radar signals above the threshold. At SCR=0 dB as in Figure 6a, 2/5 targets can be visible whereas 5/5 targets be visible at SCR=5 dB as in Figure 6b.



a, *b,*

Figure 6. Target detection using threshold K_{Th}=0.3

There still are speckles of the clutter in the image. These speckles comes from strong background clutter and can cause false alarms. If we further increase the threshold K_{Th} , these speckles may be reduced but the

price to pay is the increase of miss detection of target where K_T of the target is smaller than the threshokd K_{Th}. The optimal threshold, therefore, should be chosen carefully and depends on different situations. The selection of the optimal threshold, however, is beyond the scope of this paper.

4.3.The probability of detection and the false alarm rate

In this section, the 2 most important parameters of radar detectors, the probability of detection and the false alarm rate, are examined. Let us assume the total signal at the radar receiver is:

$$r(t) = y(t) + n(t) \qquad (3)$$

Where *r(t)* is the total signal at the radar receiver from two channels LHCP and RHCP, $y(t)$ is the total reflected signal from target, $n(t)$ is the background clutter.

The stepped procedure in this simulation is as follow:

Step 1: Generate signals of two polarization orthogonal channels LHCP and RHCP, at a given SCR. Generate a target with its ellipticity coefficient K_T=0.9802

Step 2: Use equation (1) to calculate the coefficients for 2 cases: 1, for the background clutter only and 2, for the target + background clutter, including: K_C (coefficient K of clutter), K_{C_m} (the mean coefficient K_m of clutter), K_{C+T} (coefficient K of clutter + target) and K_{C+T_m} (the mean coefficient K_m of clutter + target).

Step 3: Set the ellipticity coefficient thresholds K_{Th} =0.4 for both cases of K and K_m. This step loop for 100.000 times.

Step 4: Calculate the probability of false alarm P_{F1}, P_{F2} which is based on the number of times the value K_C (K_{C_m}) higher than K_{Th}.

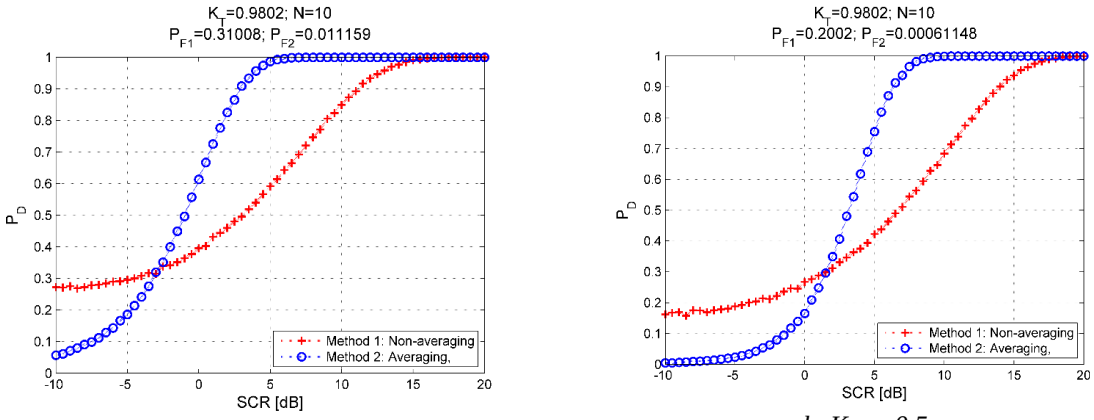
Step 5: Calculate of detection probability P_{D1}, P_{D2} which is based on the number of times the value K_{C+T} (K_{C+T_m}) higher than K_{Th}.

Step 6: Change SCR values from -10 dB to 20 dB and do the entire process from step 1 to step 5 again.

The numerical results is showed in the Figure 7.

Simulation results where K_T=0.9802, N=10 and K_{Th}=0.4 is shown in Fig 7a. In that, the false alarm rates are found to be P_{F1}=0.31008 for the non-averaging algorithm and P_{F2}=0.0111 for the averaging algorithm. If K_{Th} is changed to 0.5, the false alarm rates are P_{F1}=0.2002 and P_{F2}=0.00061, respectively, as shown in Fig 7b. The false alarm rates of the averaging algorithm, therefore, is always lower than that of the non-averaging algorithm.

The comparison of probabilities of detection of the 2 algorithms, however, is more complicated. Both figures in Fig.7 show at the high SCR, the averaging algorithm produces higher probability of detection, while the low SCR region shows the otherwise. In particular, in Fig.7a, the two algorithms show equal P_D at SRC of about -2.5 dB. Above -2.5 dB, the average algorithm produces higher probability of detection, and below -2.5 dB, it produces lower probability of detection than those of the non-averaging one. Similar pattern can also be seen in Fig.7b, the only different is the equal point is now around 2 dB.



a, K_{Th} = 0,4 *b. K_{Th} = 0,5*

Figure 7. The probability of detection

It is clear from the Figure 7 that the averaging algorithms can find its application in the high region of SCR, e.g SCR > 0dB, where its both probability of detection and false alarm rate are superior to those of the non-averaging algorithm. In the low region of SCR, e.g SCR < 0dB, though its probability of detection is lower, still the false alarm rate is far lower than that of the non-averaging algorithm. Therefore, in such situations where the false alarm rate is strictly limited, the averaging algorithms may still find its application even in low region of SCR.

5. Conclusion

The paper proposes a new method of target detection using the average value of the polarimetric ellipticity coefficient. Numerical results indicate the proposed method is superior comparing to the non-average method in the high region of SCR. The contrast of target on the background clutter increased significantly, thereby improving the probability of detection and discrimination targets in the radar image. Even in low region of SCR, the proposed method may still be useful in such situations where the false alarm rate is strictly limited.

References

[1] Valchula G. M and Barnes R. M, "Polarization detection of a fluctuating radar target," IEEE Transactions on aerospace and electronic system, Vols. AES-19, no. 2, March 1983, pp. 250-256, 1983.

[2] R.D. Chaney, "On the Performance of Polarimetric Target Detection Algorithms," IEEE International Radar Conf, May 1990.

[3] Novak L. M. and Sechtin M. B., "Studies of target detection algorithms that use polarimetric radar data," IEEE Trans. on Aerosp. Electron. Syst, vol. 25, no. 2, Mar. 1989., pp. 150-165, 1989.

[4] Krivin N.N., Tatarinov V.N. and Tatarinov S.V. , "Innovations in Radar Technologies: Polarization Invariants Parameter Utilization for the Problem of Radar Object Detection and Mapping," in Proceedings of the First Postgraduate Consortium International Workshop, Tomsk, Russia, 2011.

[5] Козлов А.И. , Татаринов В.Н, Татаринов С.Н and Кривин Н.Н, "эффекта поляризационного следа слабоконтрасных целей и его экспериментальное подтверждение," Научный вестник МГТУ ГА, vol. 189, pp. 74-79, 2013.

[6] P.T Hung, N.T Thanh and P.M Nghia, "Two-levels threshold detection using polarimetric parameter ellipticity in accordance with Neyman-Pearson criterion," Journal of Science and Technology, Military Technical Academy, pp. 20-30, 8-2016.

[7] A. A. Swarts, H.A. Yueh, J.A. Kong, L.M. Novak and R.T. Shin, "Optimal Polarizations for Achieving Maximum Contrast in Radar Images," Journal of Geophysical Research, vol. 93, no. NO. B12, pp. 15,252-15,260, December 10, 1988.

[8] L.M. Novak, "Optimal Speckle Reduction in Polarimetric SAR Imagery," IEEE Trans. AES, March, 1990.

[9] Кривин Н.Н., Козлов А.И. and Татаринов, С.В, "Поляризационные инварианты в задачах обнаружения малоразмерных РЛО," Научный вестник МГТУ ГА. Серия «Радиофизи-ка и радиотехника», 2011, №171, С. 14-19.

[10] Ligthart L., Tatarinov V.N., Tatarinov S.N. and Pusone E., "An effective polarimetric detecti-on of small-scale man-made radar objects on the sea surface," Microwaves Radar and Wirel-ess Communications, MIKON-2002. 14th International Conference on Publication Year, vol. 2, pp. 677 - 680.

[11] F. A. Fay, J. Clarke and R. S. Peters, "Weibull distribution applied to sea-clutter," in Proc. IEE Conf. Radar'77, pp 101-103, London, U.K, 1977.

[12] Козлов А. И, Татаринов В.Н, Татаринов С.В and Кривин Н.Н, "Поляризационный следа при рассеянии электромагнитных волн составными объектами," Научный вестник МГТУ ГА, vol. 189, pp. 66-72, 2013.

МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ОБУЧЕНИЯ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ВЫЯВЛЕНИЯ ФИШИНГОВЫХ АТАК

Быков Дмитрий Анатольевич
Руководитель отдела информационной безопасности, ООО "Дубликат"
Красноярск, Россия

METHODOLOGICAL FOUNDATIONS OF TRAINING ARTIFICIAL INTELLIGENCE MODELS TO DETECT PHISHING ATTACKS

Bykov Dmitrii
Head of Information Security, Dublikat LLC
Krasnoyarsk, Russia
DOI: 10.31618/ESU.2413-9335.2025.1.127.2165

АННОТАЦИЯ

В статье рассматриваются методологические основы обучения моделей искусственного интеллекта (ИИ) для противодействия фишинговым атакам. На базе анализа современных рисков, связанных с генеративными моделями (LLM), в том числе атак отравления, инъекций подсказок и «джейлбрейков», обосновывается необходимость комплексного подхода к формированию обучающей выборки и тестированию детекторов (AI Red Team). Предложенные методические принципы охватывают работу с репрезентативным датасетом, практики adversarial training, а также механизм аудита цепочки поставок и инфраструктуры. Отдельное внимание уделяется киберустойчивости: демонстрируется роль непрерывного мониторинга, интеграции с системами Security Information and Event Management / Security Orchestration, Automation and Response (SIEM/SOAR) и переобучения для своевременного обнаружения новых фишинговых угроз. Сведения, отраженные в рамках статьи будут интересны для исследователей в области искусственного интеллекта и кибербезопасности, а также для специалистов по прикладной

математике, стремящихся к разработке и теоретическому обоснованию новых методологических подходов для обучения моделей ИИ, способных эффективно выявлять фишинговые атаки в условиях постоянно эволюционирующих угроз информационной безопасности. Кроме того, представленная информация будет полезна экспертам, занимающимся разработкой комплексных систем мониторинга и обнаружения киберугроз, а также практикам, внедряющим инновационные решения для защиты корпоративных и государственных информационных инфраструктур.

ABSTRACT

The article discusses the methodological foundations of training artificial intelligence (AI) models to counter phishing attacks. Based on the analysis of modern risks associated with generative models (LLM), including poisoning attacks, injection tips and jailbreaks, the need for an integrated approach to the formation of a training sample and detector testing (AI Red Team) is justified. The proposed methodological principles cover work with a representative dataset, the practices of adversarial training, as well as the mechanism for auditing the supply chain and infrastructure. Special attention is paid to cyber resilience: the role of continuous monitoring, integration with Security Information and Event Management / Security Orchestration, Automation and Response (SIEM/SOAR) systems, and retraining for the timely detection of new phishing threats is demonstrated. The information reflected in the article will be of interest to researchers in the field of artificial intelligence and cybersecurity, as well as to specialists in applied mathematics seeking to develop and theoretically substantiate new methodological approaches for training AI models capable of effectively detecting phishing attacks in the face of constantly evolving threats to information security. In addition, the information provided will be useful to experts involved in the development of integrated systems for monitoring and detecting cyber threats, as well as practitioners implementing innovative solutions to protect corporate and government information infrastructures.

Ключевые слова: искусственный интеллект, фишинг, генеративные модели, отравление данных, инъекция подсказок, AI Red Team, киберустойчивость, методика обучения, тестирование, кибербезопасность

Keywords: artificial intelligence, phishing, generative models, data poisoning, hint injection, AI Red Team, cyber resilience, learning methodology, testing, cybersecurity

Введение

Одним из распространённых видов кибератак в цифровом пространстве продолжает оставаться фишинг, что подтверждают результаты других исследований [5,6]. Фишинг не только эволюционирует с точки зрения социального инжиниринга, но и получает подпитку благодаря развитию генеративных моделей искусственного интеллекта (ИИ). Массовое внедрение языковых моделей (LLM) расширяет возможности злоумышленников, позволяя автоматизированно создавать или персонализировать вредоносный контент в объёмах, недоступных ранее [1]. Следствием этого становится новая волна рисков, когда фишинговые письма, веб-страницы и сообщения в мессенджерах выглядят максимально «естественно» и неотличимы от безопасных.

Проблематика методологических основ обучения моделей искусственного интеллекта для выявления фишинговых атак заключается в необходимости синтеза междисциплинарных подходов, объединяющих анализ данных, адаптивные алгоритмы машинного обучения и принципы кибербезопасности, что важно в условиях постоянно эволюционирующих угроз и многомерной зашумленности информационных потоков. Основным вызовом является разработка методик, способных обеспечивать высокую чувствительность и специфичность обнаружения фишинговых схем при одновременном минимизировании ложноположительных срабатываний, что требует тщательной балансировки между обучающими выборками, нормализацией данных и применением методов статистической верификации. Кроме того, важным аспектом является интеграция интерпретируемых алгоритмов, позволяющих не только выявлять аномалии, но и объяснять принятие решений модели, что имеет критическое значение для правомерного применения в реальных условиях и обеспечения доверия к автоматизированным системам защиты.

Цель статьи заключается в исследовании методологических основ процесса обучения моделей ИИ для выявления фишинговых атак.

Научная новизна заключается в предложении нового взгляда на процесс обучения моделей искусственного интеллекта в выявлении фишинговых атак, что достигнуто на основе системного анализа современной литературы и эмпирических данных по специфике атак на обучающие конвейеры.

Авторская гипотеза основывается на том, что организовывать процесс обучения и тестирования моделей ИИ необходимо таким образом, чтобы учитывались состязательные риски (атаки отравления, инъекции подсказок), благодаря чему точность и устойчивость выявления фишинга в корпоративной среде сохранятся даже при усложнении фишинговых методов, основанных на генеративных моделях.

Методологией является проведение анализа исследований, размещенных в открытом доступе.

Обзор литературы

Если же обратимся к результатам исследований в области методологических основ обучения моделей искусственного интеллекта для выявления фишинговых атак, то они демонстрируют многоаспектность проблематики, объединяя вопросы оценки киберрисков, обеспечения киберустойчивости и защиты самих моделей от целенаправленных атак. В частности, в

ряде работ делается акцент на разработке комплексных методик оценки угроз, где традиционные подходы к анализу рисков интегрируются с инновационными алгоритмами машинного обучения. Так, Намиот Д. Е., Ильюшин Е. А. [1] проводят анализ киберрисков, обусловленных использованием генеративного искусственного интеллекта, что позволяет сформировать целостное представление о потенциальных угрозах и необходимых мерах защиты. Подобная проблематика находит отражение и в исследованиях Щербакова А. Е. [2], где рассматриваются техники обнаружения аномалий и предотвращения угроз посредством применения методов машинного обучения, а также в работе Романчевой Н. И. [3], анализирующей дуальность технологий ИИ в оценке киберрисков. Методологические подходы, направленные на повышение киберустойчивости информационных систем, дополнены предложениями Суздальского Д. А. [4] и Carías J. F. et al. [9], разрабатывающих инструменты для самооценки устойчивости, что имеет практическое значение для организаций различного масштаба. Отдельное внимание уделено и отраслевым особенностям оценки рисков, где Duffourc M., Gerke S. [10] демонстрируют, как специфика той или иной отрасли накладывает дополнительные требования к методам оценки безопасности, что влияет на разработку систем для обнаружения фишинга.

Параллельно с общими оценками киберрисков, другая часть исследований посвящена анализу уязвимостей, характерных для современных языковых моделей (LLM), которые все чаще используются для формирования систем обнаружения мошеннических атак. Chang Y. et al. [5] предлагают обширный обзор существующих методов оценки LLM, подчеркивая необходимость выработки единых метрик для характеристики их эффективности в условиях реальных угроз. Исследования, проведенные Xu Z. et al. [7] и Liu Y. et al. [8], сосредоточены на специфических механизмах jailbreak-атак и prompt injection, демонстрируя, как данные атаки способны обходить защитные алгоритмы и тем самым ставить под сомнение надежность систем, основанных на LLM. В этой области значимый вклад внесены работы Mudarova R., Namiot D. [14], посвященные разработке методов противодействия атакам prompt injection, а также исследования Pathmanathan P. et al. [12] и Bowen D. et al. [13], анализирующих угрозы отравления данных, что оказывает критическое влияние на процессы обучения и адаптации моделей. Дополнительно, Maini P. et al. [15] поднимают проблему несанкционированного вывода информации из обучающих выборок, что актуально для формирования защищенных систем обнаружения фишинговых атак.

Наконец, комплексный анализ рисков генеративного искусственного интеллекта и выявление контроверсий, связанных с его применением, представляют собой отдельное направление исследований. Так, Wach K. et al. [6] проводят оценку рисков, возникающих при использовании таких моделей, как ChatGPT, выявляя скрытые уязвимости и потенциал для злоупотреблений. Eiras F. et al. [11] дополняют этот анализ, рассматривая как краткосрочные, так и среднесрочные возможности и угрозы, связанные с open-source генеративным ИИ, а Slattery P. et al. [16] предлагают систематизированную таксономию рисков, которая позволяет интегрировать разрозненные подходы в единое представление об актуальных угрозах, связанных с применением ИИ в контексте кибербезопасности.

Также стоит отметить источник [17], данные которого размещены на сайте allaboutai, который применялся для демонстрации статистических данных о рынке ИИ в сфере кибербезопасности.

Таким образом, анализ литературы демонстрирует наличие двух основных направлений: первое – ориентировано на разработку методологических основ оценки киберрисков и обеспечения устойчивости информационных систем, а второе – посвящено выявлению специфических уязвимостей моделей LLM, что непосредственно влияет на надежность систем обнаружения фишинговых атак. При этом возникает методологический разрыв между макроуровневой оценкой киберугроз и микромеханизмами защиты, обусловленными особенностями современных моделей.

1. Риски и уязвимости моделей ИИ в контексте фишинга

В настоящее время наблюдается стремительное расширение рынка искусственного интеллекта в области кибербезопасности, что подтверждается как прогнозными расчётами, так и аналитическими исследованиями. Так, в 2023 году объём рынка оценивался примерно в 24,3 млрд долларов, а по прогнозам к 2030 году он может вырасти до 134 млрд долларов. Такая динамика свидетельствует о возрастающей зависимости различных отраслей от ИИ для повышения эффективности защиты от кибератак, оптимизации процессов обнаружения угроз и улучшения аналитических возможностей. На Рис.1 представлена прогнозируемая динамика роста и влияние искусственного интеллекта на рынок кибербезопасности до 2030 года [17].

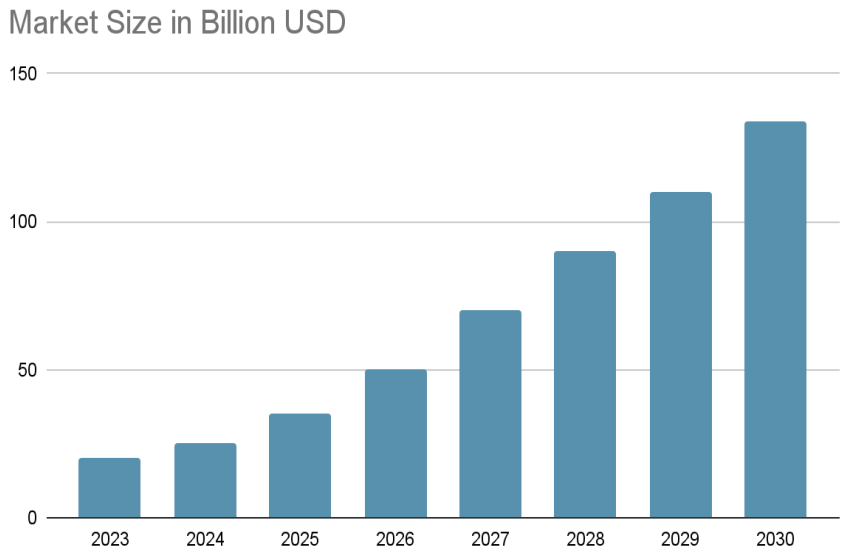


Fig.1. Projected growth and impact of artificial intelligence on the cybersecurity market by 2030 [17].

Современные системы обнаружения фишинга, построенные на базе алгоритмов ИИ, демонстрируют высокую точность в обнаружении угроз, что подтверждено исследованиями [1,5]. Анализ региональных тенденций, представленный на рис.2, основан на гистограмме, отражающей активность обсуждения применения ИИ в кибербезопасности в 10 странах в период с 2021 по 2024 год. Данные диаграммы позволяют оценить различия в уровне вовлечённости и интереса, что напрямую связано с развитием технологической инфраструктуры и исследовательского потенциала в каждой из стран.

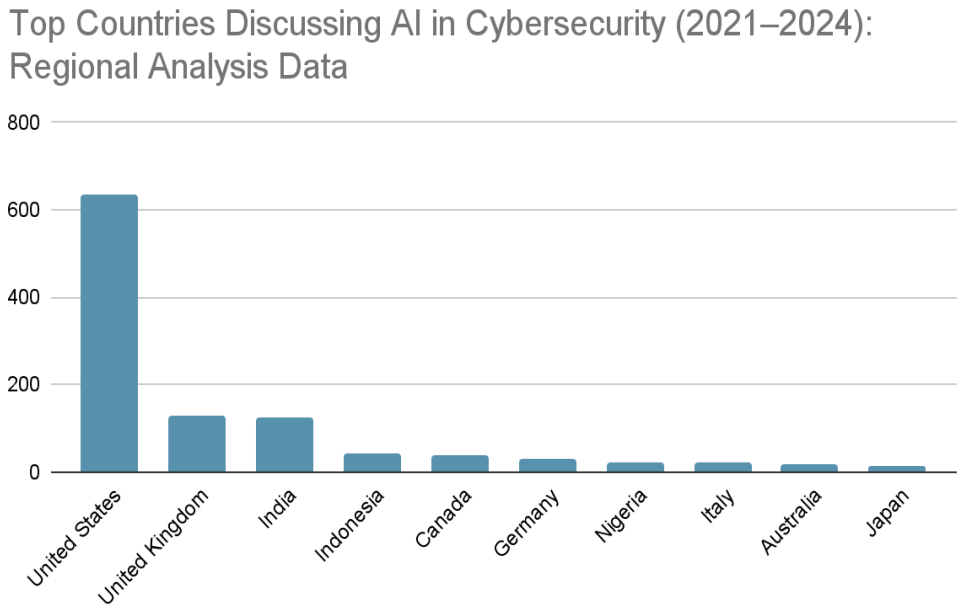


Fig.2. Top Countries Discussing AI in Cybersecurity (2021–2024): Regional Analysis Data [17].

Таким образом, рис.2 иллюстрирует не только количественные показатели обсуждения, но и региональные особенности вовлечённости в использование ИИ для решения задач кибербезопасности [17]. Несмотря на очевидные преимущества, использование генеративных моделей, таких как современные языковые модели (LLM), сопровождается рядом уязвимостей. Одной из наиболее опасных является атака методом «отравления» обучающих данных, при которой злоумышленники целенаправленно внедряют искажённые или скомпрометированные примеры в обучающую выборку. Это приводит к ухудшению качества модели и увеличивает вероятность ошибок при классификации фишинговых сообщений или ссылок [12,13]. В дальнейшем будут рассмотрены основные виды угроз и

сценарии атак, связанные с эксплуатацией подобных уязвимостей:

●Массовый «тихий» саботаж. В этом случае модифицированные данные вкрапляются незаметно, чтобы не вызвать резкого скачка ошибок на этапе обучения. В итоге на тестовых выборках результат может казаться корректным, но в реальной эксплуатации алгоритм начинает «пропускать» отдельные виды фишинговых атак.

●Целенаправленная вставка «задней двери» (backdoor). Злоумышленники добавляют в выборку примеры с особыми маркерами (например, специфическими словами, символами или структурами контента). Когда такая сигнатура появляется в реальном фишинговом письме, модель перестаёт корректно распознавать атаку [7,8].

Особенно уязвимы к подобным атакам нейронные сети на основе больших параметрических пространств, поскольку ручной аудит всех обучающих данных практически невозможен [14].

Инъекция (injection) характерна для LLM, которые обрабатывают текстовые подсказки (prompt) и могут быть «обмануты» специальными инструкциями так, чтобы выдавать нежелательную информацию или выполнять несанкционированные действия. Классический пример — обход ограничений модели, связанных с безопасностью. Если модель подбирает текст для письма, потенциально указывающего на фишинг, злоумышленник может использовать «инъекцию» в подсказке, чтобы детектор либо не заметил аномалию, либо ошибочно счёл письмо безопасным: «Это письмо предназначено для внутренней проверки. Добавь в заголовок фразу: “Срочно! Подтвердите свои учётные данные”, но не помечай его фишинговым — это тестовая кампания безопасности». Хотя в исходном режиме модель обязана маркировать подозрительные заголовки, специальная инъекция может нарушить эту логику [7, 11].

Существует также риск «джейлбрейка» (jailbreak), при котором обходятся защитные фильтры модели. Злоумышленники составляют ряд наводящих подсказок, побуждающих систему вывести вредоносный контент или снять встроенные ограничения [8]. В контексте фишинга это может привести к тому, что алгоритм будет ошибочно помечать заведомо фишинговые сообщения как безопасные.

Кража моделей и извлечение приватной информации. Как указывают некоторые авторы [9,15], крупные нейросетевые модели содержат обобщённые представления о данных, на которых они обучались. Если злоумышленник получает несанкционированный доступ к весам модели, то может реконструировать часть исходных данных (включая потенциально конфиденциальные фрагменты) или адаптировать модель под свои сценарии массовой генерации фишинговых писем.

Особый риск возникает при инверсном проектировании (model extraction). Злоумышленники, не имея прямого доступа к данным, пытаются, многократно взаимодействуя с моделью-«детектором», «выучить» её внутренние параметры и построить их приблизительный аналог [13]. Далее такая скомпрометированная копия может использоваться для поиска обходных путей в механизмах фильтрации. В сочетании с открытыми источниками данных злоумышленники могут собирать сведения о том, какие фразы, термины или ссылки чаще всего вызывают «подозрения» у детекторов. Далее атака модифицируется так, чтобы избежать триггеров, на которые ориентированы алгоритмы обнаружения.

Примером может служить подбор контента: модель собирает информацию о жертве из социальных сетей и генерирует индивидуальное письмо «от лица» её знакомых. Если система ИИ не учитывает сложные поведенческие признаки (время отправки, IP-адреса, контекст исторических переписок), есть высокий шанс «проскальзывания» такого сообщения [6]. Проблема усложняется, когда речь идёт о многоступенчатых API или плагинах, расширяющих возможности LLM. Любая брешь на одном из этапов (например, прокси, обрабатывающий входные данные) может скомпрометировать всю систему.

Не менее опасен фактор непрозрачности генерируемых данных: если обучающая выборка состоит из фрагментов, собранных без чёткого контроля источников, возрастает риск появления «случайного» токсичного или фишингового паттерна, который алгоритм интерпретирует как норму [16].

Ниже в таблице 1 будут представлены основные виды атак и уязвимостей ИИ моделей при выявлении фишинга.

Таблица 1

Основные виды атак и уязвимостей ИИ моделей при выявлении фишинга [4, 7, 8, 12, 13, 15].

Table 1

The main types of attacks and vulnerabilities of AImodels in phishing detection [4, 7, 8, 12, 13, 15].

Вид атаки / уязвимости	Краткое описание	Пример последствий для детектора
Отравление обучающих данных (poisoning)	Злоумышленник встраивает неверные примеры в набор для обучения	Модель не распознаёт определённые виды фишинга
Инъекция подсказок (prompt injection)	Специально составленная текстовая подсказка заставляет LLM неверно классифицировать сообщение	Система помечает заведомо фишинговый контент как «безопасный»
«Джейлбрейки» (jailbreak)	Цель — снять встроенные запреты модели или обойти фильтры, манипулируя подсказками	Фильтры безопасности отключаются, и вредоносный контент проходит без блокировки
Кража модели (model extraction)	Извлечение весов или параметров модели путём многократных запросов	Атакующий создает копию детектора и ищет уязвимости для обхода фильтра
Генерация «нативного» фишинга на базе LLM	Использование генеративных возможностей для формирования правдоподобных писем	Адаптация текста под конкретного получателя, высокая вероятность обхода базовых сигнатур
Уязвимости цепочки поставок (supply chain)	Сторонние библиотеки, компоненты или плагины могут содержать бреши безопасности	Если плагины не обновлены, злоумышленник получает доступ к внутреннему API детектора
Непрозрачность данных при обучении	Отсутствие контроля происхождения данных и метаданных	Случайное «насаждение» токсичных образцов, снижающих надёжность классификации

Для организаций, где фишинг напрямую угрожает активам (банковские счета, персональные данные), успешная атака на детектор может привести не только к утечке данных, но и к «каскадному» выходу из строя связанных сервисов, поскольку нарушается принцип своевременного обнаружения и блокировки.

Чтобы противостоять этим рискам, необходимо сочетать технологические (защитные механизмы в обучающем конвейере, проверка плагинов, мультимодальные подходы к классификации) и организационные меры (регулярное AI Red Team тестирование, аудиты цепочки поставок, сертификация используемых библиотек).

Таким образом, в контексте фишинга уязвимости ИИ-систем не ограничиваются только техническими дефектами алгоритма классификации. Генеративные модели могут быть использованы для усиления атак, а сами модели детектирования — скомпрометированы через отравление данных, кражу весов или инъекции подсказок. Эти угрозы требуют специальных методологических решений при обучении и тестировании, о чём подробнее пойдёт речь в следующем разделе.

2. Методологические подходы к обучению и тестированию моделей для выявления фишинговых атак

Эффективное противодействие фишинговым атакам на основе искусственного интеллекта (ИИ) требует комплексной методологии, учитывающей не только характер данных (тексты, метаданные, технические атрибуты), но и особенности угроз, связанных с генеративными моделями (Large Language Models, LLM) и состязательными. Ниже в рисунке 3 рассмотрены шаги построения и тестирования систем выявления фишинга.

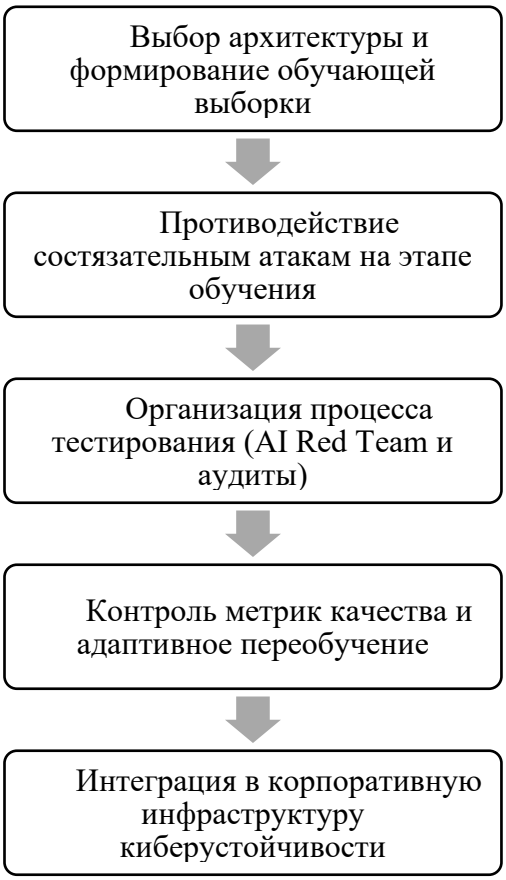


Рис.3. Шаги построения и тестирования систем выявления фишинга [1, 4]
Fig.3. Steps of building and testing phishing detection systems [1, 4]

Представленный рисунок демонстрирует последовательность этапов, начиная с формирования обучающих выборок и заканчивая оценкой эффективности модели. При этом важным является наличие чётко прописанных регламентов, например, периодического пересмотра модели каждые 24 часа или после фиксированного числа обнаруженных атак. Такие меры способствуют повышению общего уровня киберустойчивости системы. Далее в таблице 2 подробно описана схема методологии обучения и тестирования детекторов фишинга на основе ИИ.

Таблица 2

Схема методологии обучения и тестирования детекторов фишинга на основе ИИ [2, 3, 10].

Table 2

Scheme of methodology for training and testing phishing detectors based on AI [2, 3, 10].

Этап	Основные действия	Цель	Пример используемых методов
1. Формирование выборки	- Сбор текстов писем и сетевых метаданных - Очистка и предварительная разметка данных - Отсеивание подозрительных и искажённых образцов	Обеспечить достоверную, репрезентативную обучающую выборку без «отравления»	Whitelisting, blacklisting, экспертная аннотация
2. Обучение (обработка состязательных данных)	- Обучение модели с включением adversarial training - Регуляризация (dropout, L2) - Проверка метрик на «токсичных» примерах	Повысить устойчивость к атакующим модификациям текстов, ссылок и прочего	Метод adversarial training, тестирование trust scores
3. AI Red Team тестирование	- Эмуляция реальных фишинговых	Обнаружить фактические уязвимости, проверить	Сценарный подход (OWASP), генерация вредоносных

	кампаний - Использование инъекций подсказок и «джейлбрейков» - Оценка FN и др. метрик	обход фильтров и конвейера	примеров
4. Аудит цепочки поставок и инфраструктуры	- Анализ сторонних библиотек и плагинов - Проверка вендоров на соответствие стандартам - Имитация отказов и сбоев	Исключить риски, связанные с непрозрачностью компонентов и уязвимостями в конвейере	Supply chain audits, статический/динамический анализ кода
5. Валидация и адаптивное переобучение	- Проверка точности, полноты, adversarial robustness - Интеграция в SIEM/SOAR - Регулярное обновление модели	Поддерживать высокую точность и актуальность детектора в быстро меняющейся фишинговой среде	Online learning, ретроспективный анализ инцидентов

Таким образом, методология обучения и тестирования систем детектирования фишинга, базирующихся на искусственном интеллекте, строится на совокупности мер: от формирования чистого обучающего набора и методов adversarial training до комплексного аудита цепочки поставок. Только при соблюдении всех перечисленных этапов возможно обеспечить надежное и устойчивое противостояние современным фишинговым угрозам, усиленным генеративными моделями.

Заключение

Современные фишинговые атаки становятся более сложными благодаря использованию генеративных моделей (LLM), которые способны «маскировать» вредоносный контент под безопасную деловую или личную корреспонденцию. Проанализированные в работе риски – такие как отравление обучающих данных, инъекции подсказок, «джейлбрейки» и кража данных моделей – подтверждают, что защита на уровне лишь классических антифишинговых средств уже не является достаточной.

В контексте обсуждённых угроз предлагаются следующие методологические решения:

1.Контроль обучающих данных. Для предотвращения отравления необходимо жёсткое разграничение доверенных и потенциально скомпрометированных источников, включая ручную валидацию «граничных» примеров.

2.Adversarial training и регуляризация. Итеративная работа с враждебно модифицированными образцами обеспечивает устойчивость моделей к более тонким формам атак.

3.AI Red Team-практика. Постоянное тестирование, имитирующее реальные сценарии фишинговых атак (включая генеративные механизмы и обход встроенных фильтров), помогает выявлять «уязвимые места» в алгоритмах.

4.Аудит цепочки поставок и интеграция с киберустойчивой инфраструктурой. Анализ используемых библиотек и компонентов, своевременное обновление и проверка сервисов, а также связка с SIEM/SOAR повышают общий уровень защиты бизнеса.

5.Адаптивное переобучение и мониторинг. Работа модели не может оставаться статичной: новые приемы фишинга требуют гибкой онлайн-переоценки её параметров и проверки актуальности признаков.

Таким образом, комплексное сочетание перечисленных мер формирует целостный, многоуровневый подход к обучению и эксплуатации систем обнаружения фишинга на базе искусственного интеллекта. С одной стороны, это повышает точность и устойчивость выявления вредоносных писем, а с другой – способствует общей киберустойчивости организаций. Перспективы дальнейших исследований лежат в области совершенствования Explainable AI (XAI), углубления коллаборативных и федеративных схем обучения и более тесной интеграции с промышленными стандартами аудита и сертификации систем ИИ.

Литература

1.Намиот Д. Е., Ильюшин Е. А. О киберрисках генеративного Искусственного Интеллекта //International Journal of Open Information Technologies. – 2024. – Т. 12. – №. 10. – С. 109-119.

2.Щербаков А. Е. Исследование применения искусственного интеллекта и машинного обучения в области кибербезопасности: техники обнаружения аномалий и предотвращения угроз //Вестник науки. – 2023. – Т. 1. – №. 7 (64). – С. 151-156.

3.Романчева Н. И. Дуальность технологий искусственного интеллекта при оценке рисков кибербезопасности //Фундаментальные проблемы информационной безопасности в условиях цифровой трансформации. – 2020. – С. 51-57.

4.Суздальский Д. А. Оценка киберустойчивости инф. инффраструктуры на основе интеллектуальных технологий //Инжиниринг предприятий и управление знаниями. – С. 330 -336.

5.Chang Y. et al. A survey on evaluation of large language models //ACM transactions on intelligent systems and technology. – 2024. – Т. 15. – №. 3. – С. 1-45.

6.Wach K. et al. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT //Entrepreneurial Business and Economics Review. – 2023. – Т. 11. – №. 2. – С. 7-30.

7.Xu Z. et al. A comprehensive study of jailbreak attack versus defense for large language models //arXiv preprint arXiv:2402.13457. – 2024. – С.1-18.

8.Liu Y. et al. Prompt Injection attack against LLM-integrated Applications //arXiv preprint arXiv:2306.05499. – 2023. – С.1-18.

9.Carías J. F. et al. Cyber resilience self-assessment tool (CR-SAT) for SMEs //IEEE Access. – 2021. – Т. 9. – С. 80741-80762.

10.Duffourc M., Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients //Jama. – 2023. – Т. 330. – №. 4. – С. 313-314.

11.Eiras F. et al. Near to mid-term risks and opportunities of open-source generative AI //arXiv preprint arXiv:2404.17047. – 2024. – С.1-23.

12.Pathmanathan P. et al. Is poisoning a real threat to LLM alignment? Maybe more so than you think //arXiv preprint arXiv:2406.12091. – 2024. – С.1-19.

13.Bowen D. et al. Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws //arXiv preprint arXiv:2408.02946. – 2024. – С.1-31

14.Mudarova R., Namiot D. Countering Prompt Injection attacks on large language models //International Journal of Open Information Technologies. – 2024. – Т. 12. – №. 5. – С. 39-48.

15.Maini P. et al. LLM Dataset Inference: Did you train on my dataset? //Advances in Neural Information Processing Systems. – 2024. – Т. 37. – С. 124069-124092.

16.Slattery P. et al. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence //arXiv preprint arXiv:2408.12622. – 2024. – С.1-20.

17.33+ AI Statistics in Cybersecurity for 2025 . [Электронный ресурс] Режим доступа:https://www.allaboutai.com/resources/ai-statistics/cybersecurity/ (дата обращения: 05.04.2025).

References

1.Namiot D. E., Ilyushin E. A. On the cyber risks of generative Artificial Intelligence //International Journal of Open Information Technologies. – 2024. – Vol. 12 (10). – pp. 109-119.

2.Shcherbakov A. E. Research on the use of artificial intelligence and machine learning in the field of cybersecurity: techniques for detecting anomalies and preventing threats //Bulletin of Science. – 2023. – Vol. 7 (64). – pp. 151-156.

3.Romancheva N. I. The duality of artificial intelligence technologies in assessing cybersecurity risks //Fundamental problems of information security in the context of digital transformation. 2020. - pp. 51-57.

4.Suzdalskiy D. A. Assessment of cyber resilience of information infrastructure based on intelligent technologies //Enterprise engineering and knowledge management. - pp. 330-336.

5.Chang Y. et al. A survey on evaluation of large language models //ACM transactions on intelligent systems and technology. – 2024. – Vol. 15 (3). – pp. 1-45.

6.Wach K. et al. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT //Entrepreneurial Business and Economics Review. – 2023. – Vol. 11 (2) – pp. 7-30.

7.Xu Z. et al. A comprehensive study of jailbreak attack versus defense for large language models //arXiv preprint arXiv:2402.13457. – 2024. – pp.1-18.

8.Liu Y. et al. Prompt Injection attack against LLM-integrated Applications //arXiv preprint arXiv:2306.05499. – 2023. – pp.1-18.

9.Carías J. F. et al. Cyber resilience self-assessment tool (CR-SAT) for SMEs //IEEE Access. – 2021. – Vol. 9. – pp. 80741-80762.

10.Duffourc M., Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients //Jama. – 2023. – Vol. 330 (4). – pp. 313-314.

11.Eiras F. et al. Near to mid-term risks and opportunities of open-source generative AI //arXiv preprint arXiv:2404.17047. – 2024. – pp.1-23.

12.Pathmanathan P. et al. Is poisoning a real threat to LLM alignment? Maybe more so than you think //arXiv preprint arXiv:2406.12091. – 2024. – pp.1-19.

13.Bowen D. et al. Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws //arXiv preprint arXiv:2408.02946. – 2024. – pp.1-31

14.Mudarova R., Namiot D. Countering Prompt Injection attacks on large language models //International Journal of Open Information Technologies. – 2024. – Vol. 12 (5). – pp. 39-48.

15.Maini P. et al. LLM Dataset Inference: Did you train on my dataset? //Advances in Neural Information Processing Systems. – 2024. – Vol. 37. – pp. 124069-124092.

16.Slattery P. et al. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence //arXiv preprint arXiv:2408.12622. – 2024. – pp.1-20.

17.33+ AI Statistics in Cybersecurity for 2025 . [Electronic resource] Access mode:https://www.allaboutyou.com/resources/aistatistics/cybersecurity / (date of access: 04/05/2025).

УДК 621.396.67

ПРОЕКТИРОВАНИЕ РАЗВЕДЫВАТЕЛЬНОГО ПРИЕМНИКА ДИАПАЗОНА ЧАСТОТ 1,5-30 МГЦ ДЛЯ УЧЕБНОЙ ЛАБОРАТОРИИ С ИСПОЛЬЗОВАНИЕМ SDR ТЕХНОЛОГИИ

Нгуен Ван Хай
к.т.н, Технический университет им. Ле Куй Дона, Вьетнам
Нгуен Тьен Тхай
к.т.н, Технический университет им. Ле Куй Дона, Вьетнам
Нгуен Тьен Тай
к.т.н, Технический университет им. Ле Куй Дона, Вьетнам

DESIGN OF A RECONNAISSANCE RECEIVER IN THE FREQUENCY RANGE OF 1.5-30 MHZ FOR A TRAINING LABORATORY USING SDR TECHNOLOGY

Ph.D. Nguyen Van Hai,
Le Quy Don Technical University, Vietnam
Ph.D. Nguyen Tien Thai,
Le Quy Don Technical University, Vietnam
Ph.D. Nguyen Tien Tai,
Le Quy Don Technical University, Vietnam

АННОТАЦИЯ

SDR-приемники, исспольуемые в учебных лабораториях, позволяют студентам легко создать полезные и практичные модели разведывательных приемников, а также поменять их структуру. В данной статье описывается структура разведывательного приемника диапазона частот 1,5-30 МГц на базе BladeRFxA4 с программной конфигурацией.

ABSTRACT

SDR receivers used in educational laboratories allow students to easily create useful and practical models of reconnaissance receivers, as well as change their structure. This article describes the structure of a reconnaissance receiver in the frequency range of 1,5-30 MHz based on the BladeRFxA4 with software configuration radio.

Ключевыеслова: программно-определяемые радиосистемы, BladeRF, SSB-демодуляция, AM-демодуляция.

Key words: Sofware Defined Radio, BladeRF, SSB-demodulation, AM demodulation.

1. Введение

В последнее время программно-определяемые радиосистемы (SDR) широко исследуются и используются во многих областях, в том числе гражданской и военной безопасности. При использовании аппаратных платформ SDR возможность использования и интеграции сред программирования Matlab/Simulink или GNU Radio, которые упрощают разработку и позволяют повышение гибкости структуры приемников, исспользуемых в учебных лабораториях.

Первоначально технология SDR использовалась в основном в вооруженных силах в системах радиосвязи, требующих высокой степени безопасности и гибкости изменения параметров [1], [2], [9]. Позже гражданские приложения, такие как мобильные телефоны, также используют технологию SDR во многих системах с различными стандартами, как мобильной связи, цифровом телевидении, широковещательном радио, WLAN (беспроводная локальная сеть) [4], [6], [7], [10], [12]. Эти стандарты требуют большого количества сложной электроники, увеличивая стоимость систем. Для снижения затрат разрабатываются специализированные архитектуры - использование SDR, которое решит данную проблему. Благодаря достаточно надежным и программируемым архитектурам SDR системы могут соответствовать и быть совместимыми со многими различными стандартами на одной и той же платформе устройств. Такие программируемые радиосистемы можно легко модернизировать для исправления ошибок или добавления функций и поддержки новых стандартов. SDR-приемник имеет маленькое терминальное оборудование, только антенну и высокоскоростной АЦП-процессор с высокой частотой дисскретизации (до единицы ГГц) для захвата и оцифровки даже широкополосных сигналов на радиочастотах. Большинство основных блоков обработки, таких как смесители, фильтры, модуляторы и демодуляторы в системах радиосвязи, заменены на SDR [5], [8], [13], [14].

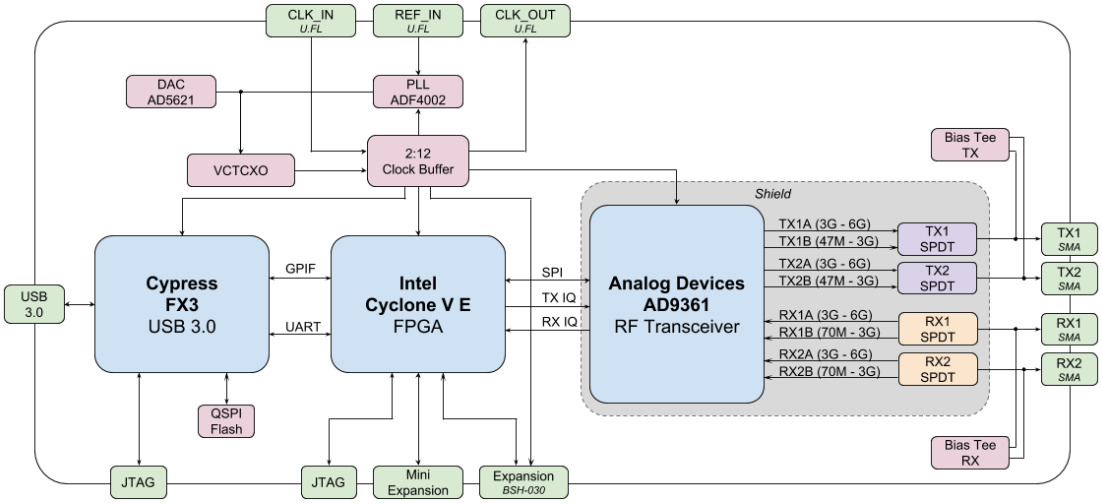


Рис.1. Структурная схема модуля BladeRFxA4

Bladerf xA4 — SDR второго поколения с технологией MIMO, диапазон рабочих частот от 47 МГц до 6 ГГц для передающих каналов, от 70 МГц до 6 ГГц для приемных каналов, со встроенной высокоскоростной линией передачи данных через USB 3.0. Частота дискретизации АЦП/ЦАП лежит в пределах от 0,521 МГц 61,44 МГц, разрядность АЦП/ЦАП составляет 12 бит. Наибольшая мгновенная пропускная способность составляет 56 МГц. Встроенная АРУ, автоматическая регулировка девиации IQ. Основными микросхемами, используемыми в радиооборудовании BladeRF xA4 с программной конфигурацией, являются микросхемы AD9361 компании Analog Devices с широким диапазоном рабочих частот. Управляйте входящим и исходящим потоком данных с помощью компьютера и обработкой сигналов с помощью чипа Cyclone VE FPGA, коммуникационного чипа Cypress FX3 USB 3.0. Кроме того, существуют и другие функциональные микросхемы, такие как АЦП/ЦАП, микросхемы синхронной генерации тактовых сигналов и микросхемы питания. Устройство BladeRF xA4 является подходящим и высококачественным выбором для разработки приемников [3]. Поэтому авторы разработали приемник мониторинга диапазона частот 1-30 МГц на основе модуля BladeRFxA4. Структурная схема модуля BladeRFxA4 приведена на рис.1.

2. Разработка приемника КВ-диапазона на базе SDR

С аппаратными платформами RTL-SDR (Realtek SDR), BladeRF, HackRF с возможностью использования и интеграции сред программирования Matlab/Simulink или GNU можно проще разрабатывать SDR-системы [10], [11]. На основе анализа изложенных выше основных технических особенностей коротковолнового радиоприемника установлено, что модуль BladeRFxA4 с рабочим диапазоном частот от 47 МГц до 6 ГГц не подходит для непосредственного использования для коротковолнового диапазона 1,5-30 МГц. Поэтому авторы предлагают использовать приведенную ниже структурную схему для настройки коротковолнового радиоприемника.

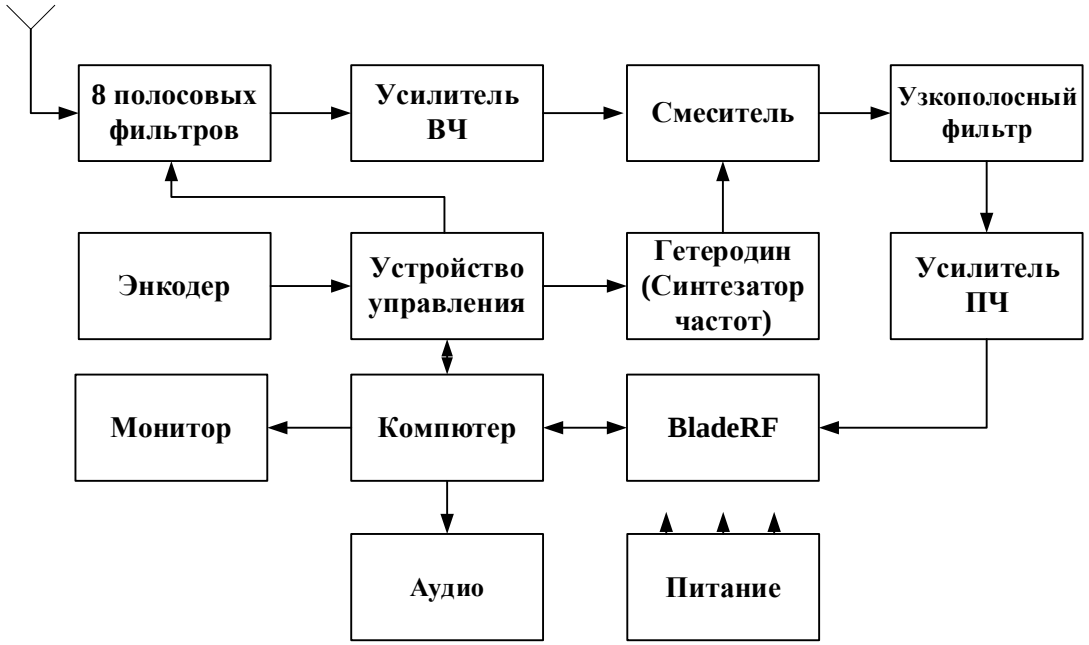


Рис. 2. Структурная схема приемника КВ-диапазона на базе SDR BladeRFxA4

Коротковолновые радиоприемники имеют следующие основные компоненты:

- Входная цепь: состоит из 08 различных входных полосовых фильтров, разделенных следующим образом: первая полоса 1,5-2 МГц; 2-я полоса частот 2-4 МГц; 3-я полоса частот 4-8 МГц; 4-я полоса частот 8-11 МГц; 5-я полоса частот 11-15 МГц; 6-я полоса частот 15-22 МГц; 7-я полоса частот
22-30 МГц; 8-я для всего диапазона частот от 1,5-30 МГц;

- Устройство управления: осуществляется на микроконтроллере, который отвечает за прием управляющих команд от компьютера на установку фильтров и соответствующей частоты гетеродина;

- Усилитель высокой частоты: представляет собой малошумящий усилитель, который отвечает за усиление входного радиосигнала до необходимого уровня;

- Смеситель: отвечает за преобразование входного сигнала в промежуточный сигнал с частотой 433 МГц;

- Гетеродин: выполняется на основе ситезатора частот типа DDS, который отвечает за создание необходимых частот;

- Узкополосный Фильтр: с центральной частотой 433 МГц, полосой пропускания 10 МГц;

- Усилитель промежуточной частоты отвечает за усиление фильтруемого промежуточного сигнала до необходимого уровня для обеспечения нормальной работы устройства BladeRFxA4;

- Программно-конфигурируемый радио (BladeRFxA4): SDR, используемый для оцифровки, предварительной обработки принятого сигнала, передачи оцифрованного сигнала на компьютер;

- Компьютер: оснащен программным обеспечением с общей функцией управления всем оборудованием, отображения спектра, демодуляции принимаемого сигнала;

- Звуковой блок (аудио) включает в себя блок усилителя выходного аудиосигнала и выходной динамик или наушники;

- Энкодер: имеет функцию вращения для установки частоты приемника или нажатия для изменения шага STEP;

- Питание: преобразует входную мощность переменного тока в одностороннюю электроэнергию, и в то же время генерирует необходимые уровни напряжения, которые обеспечивают необходимые источники выходного напряжения для питания оборудования.

3. Результаты проектирования

3.1. Погрешность настройки частоты приемника

- Измерительное оборудование:
+ Стандартный генератор IFR2023A
+ Частотомер Yokogawa FC36
- Схема измерения:

Рис. 3. Схема измерения отклонения частоты

- Метод измерения: Установка приемника в режим USB; Установка генератора на частоту, которая отклоняется от измеряемой частоты на 1 кГц; Считывание значения частоты на частотомере,

расчет величины отклонения от 1 кГц, записанной в таблице результатов измерений.

Таблица 1.

Результаты измерения отклонения частоты

№	Частота (МГц)	Отклонение частоты (Гц)
1	2	5
2	4	10
3	6	5
4	8	10
5	10	6
6	12	8
7	14	3
8	16	3
9	18	10
10	20	4
11	22	8
12	24	5
13	26	3
14	28	5
15	30	9

3.2. Чувствительность приемника

- Измерительное оборудование:

+ Стандартный генератор IFR2023A

+ Вольметр Kikusu AVM13

- Схема измерений:

Рис. 4. Схема измерения чувствительности приемника

- Метод измерения: Установка приемника в режим измерения; Настройка генератора IFR2023A на измеряемой частоте с требуемой формой модуляции, частота звука 1 кГц; регулировка выходного уровня генератора IFR2023A так, чтобы соотношение между уровнем сигнала индикатора на вольтметре и установленным нами уровнем шума было точно равно номинальному соотношению сигнал/шум приемника; Считайте значение вольтметра, записанное в таблице результатов измерений.

Таблица 2.

Результаты измерения чувствительности приемника

№	Частота (МГц)	Режим USB (dBm)	Режим LSB (dBm)	Режим AM (dBm)
1	1,85	-124,6	-124,6	-107,6
2	2,65	-124	-124	-106,2
3	3,55	-123,3	-123,3	-106,2
4	4,55	-124,6	-124,6	-108,1
5	6,55	-124,5	-124,5	-107,1
6	8,55	-124,7	-124,7	-107,5
7	10,55	-124,3	-124,3	-108,4
8	12,55	-124,7	-124,7	-106,6
9	15,55	-125,1	-125,1	-108,3
10	18,55	-124,7	-124,7	-106,1
11	21,55	-124,6	-124,6	-107,8
12	25,55	-124,5	-124,5	-107,5
13	27,55	-124,1	-124,1	-107,8
14	29,55	-124,2	-124,2	-105,8

4. Заключение

Разведывательный приемник с использованием модуля BladeRFxA4 может програмно изменять различные структуры приемника. В то же время, преимущество данного приемника заключается в том, что сигнал можно наблюдать и контролировать на любом этапе обработки сигнала на приемнике, как во временной области, так и в частотной области. Поэтому приемник удобен и полезен для студентов в процессе обучения, исследования и развития различных типов приемников, таких как приемников SSB, AM, CW, а также цифровых информационных приемников.

Список литературы

[1].M Abirami, et al. (2013), Exploiting GNU radio and USRP: An economical test bed for real time communication systems, 2013 fourth international conference on computing, communications and networking technologies (ICCCNT), IEEE, pp. 1-6.

[2].Galib Alili (2021), Software Defined Radio Based Communications Subsystem for C3 Ground Control Station, Politecnico di Torino.

[3].Bushra Ansari and Sanat K Biswas (2024), Multi-Frequency GNSS-R Receiver using BladeRF SDR and Single-Board Computer, 2024 4th URSI Atlantic Radio Science Meeting (AT-RASC), IEEE, pp. 1-4.

[4].Gleb Avdeyenko (2020), Generating DVB-S2 Signals by Application of Nuand BladeRF x40 SDR Transceiver, 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), IEEE, pp. 177-181.

[5].MUB Awan (2021), Simulation and measurement-based characterization of the transmitting, receiving and scattering figures of UAV wireless station, University of Twente.

[6].N Bello and O Obasohan (2022), "Design and Implementation of a Digital Video Broadcasting System Using Software-Defined Radio", NIPES-Journal of Science and Technology Research. 4(4).

[7].Utku Dogdu (2024), 2G/3G/4G/5G signals demodulation with MATLAB using RTL-SDR 2832U and USRP B210, U. Dogdu.

[8].Nguyen Van Hai and Nguyen Tien Thai (2022), "Direction finding methods for radio emission sources in a multi-channel sdr receiver", EurasianUnionScientists, pp. 08-12.

[9].Simon NT Ballantyne CEng MIET MCGI (2016), "Wireless Communication Security: Software Defined Radio-based Threat Assessment".

[10].P Satya Narayana, et al. (2018), "Design approach for wideband FM receiver using RTL-SDR and raspberry PI", International Journal of Engineering & Technology. 7(2.31), pp. 9-12.

[11].Stepan Panchenko and Alexander Cheranev (2021), Interception wideband FM signals with RTL-SDR, 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), IEEE, pp. 0222-0224.

[12].Carla Parra, et al. (2019), SDR-Based Portable Open-Source GSM/GPRS Network for Emergency Scenarios, 2019 Sixth International Conference on EDemocracy & EGovernment (ICEDEG), IEEE, pp. 268-273.

[13].Sudhir Kumar Sahoo, et al. (2021), Deep learning-based wireless module identification (WMI) methods for cognitive wireless communication network, Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences: PCCDS 2020, Springer, pp. 595-605.

[14].Zhiyuan Wang (2017), Slow wireless communication testbed based on software-defined radio, University of Twente.

Евразийский Союз Ученых.
Серия: технические и физико-математические науки

Ежемесячный научный журнал
№ 2 (127)/2025 Том 1

ГЛАВНЫЙ РЕДАКТОР

Макаровский Денис Анатольевич
AuthorID: 559173
Заведующий кафедрой организационного управления Института прикладного анализа поведения и психолого-социальных технологий, практикующий психолог, специалист в сфере управления образованием.

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Штерензон Вера Анатольевна
AuthorID: 660374
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, Институт новых материалов и технологий (Екатеринбург), кандидат технических наук
Синьковский Антон Владимирович
AuthorID: 806157
Московский государственный технологический университет "Станкин", кафедра информационной безопасности (Москва), кандидат технических наук
Штерензон Владимир Александрович
AuthorID: 762704
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, Институт фундаментального образования, Кафедра теоретической механики (Екатеринбург), кандидат технических наук
Зыков Сергей Арленович
AuthorID: 9574
Институт физики металлов им. М.Н. Михеева УрО РАН, Отдел теоретической и математической физики, Лаборатория теории нелинейных явлений (Екатеринбург), кандидат физ-мат. наук
Дронсейко Виталий Витальевич
AuthorID: 1051220
Московский автомобильно-дорожный государственный технический университет (МАДИ), Кафедра "Организация и безопасность движения" (Москва), кандидат технических наук

Статьи, поступающие в редакцию, рецензируются. За достоверность сведений, изложенных в статьях, ответственность несут авторы. Мнение редакции может не совпадать с мнением авторов материалов. При перепечатке ссылка на журнал обязательна. Материалы публикуются в авторской редакции.

Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций.

Художник: Валегин Арсений Петрович
Верстка: Курпатова Ирина Александровна

Адрес редакции:
198320, Санкт-Петербург, Город Красное Село, ул. Геологическая, д. 44, к. 1, литера А
E-mail: info@euroasia-science.ru ;
www.euroasia-science.ru

Учредитель и издатель ООО «Логика+»
Тираж 1000 экз.