ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ
ФАКУЛТЕТ ПО ТЕЛЕКОМУНИКАЦИИ
КАТЕДРА „КОМУНИКАЦИОННИ МРЕЖИ“

маг. инж. Атанас Георгиев Влахов

ИНТЕЛИГЕНТНО УПРАВЛЕНИЕ НА МРЕЖИ ЗА ДОСТЪП С ОТВОРЕНИ ИНТЕРФЕЙСИ ЗА РЕАЛИЗАЦИЯ НА УСЛУГИ, КРИТИЧНИ КЪМ КАЧЕСТВОТО НА ОБСЛУЖВАНЕ

АВТОРЕФЕРАТ

на дисертация за присъждане на образователна и научна степен

„ДОКТОР”

Област на висше образование: 5. Технически науки

Професионално направление: 5.3 Комуникационна и компютърна техника

Научна специалност: Комуникационни мрежи и системи

Научен ръководител: проф. дтн инж. Владимир Костадинов Пулков

София, 2026

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Комуникационни Мрежи“ към Факултет по Телекомуникации при Технически Университет – София на редовно заседание, проведено на 14.04.2026 г. (протокол № 9)

Публичната защита на дисертационния труд ще се състои на 07.07.2026г. от 15:00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед № ОЖ-5.3-36 / 04.05.2026 г. на Ректора на ТУ-София в състав:

1. проф. д-р Георги Илиев – председател
2. доц. д-р Георги Балабанов – научен секретар
3. проф. д-р Станимир Садинов
4. проф. д-р Розалина Димова
5. проф. д-р Габриела Атанасова

Рецензенти:

1. проф. д-р Георги Илиев
2. проф. д-р Габриела Атанасова

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет по Телекомуникации, блок 1, стая 1439-Б и на Интернет страницата на Технически Университет – София.

Дисертантът е задочен докторант към катедра „Комуникационни Мрежи“, Факултет по Телекомуникации. Изследванията по дисертационният труд са направени от автора, като резултатите от тях са публикувани.

Автор: маг. инж. Атанас Георгиев Влахов

Заглавие: Интелигентно управление на мрежи за достъп с отворени интерфейси за реализация на услуги, критични към качеството на обслужване

Тираж: 30 броя

Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

В последното десетилетие телекомуникационната индустрия преминава през фундаментална трансформация, обусловена от експоненциалния ръст на мобилния трафик и появата на нови, услуги. Навлизането на петото (5G) и подготовката за шестото (6G) поколение мобилни мрежи изисква поддръжка на разнородни сценарии на употреба - от масова комуникация между машини (mMTC) до свръхнадеждна комуникация с ниска латентност (URLLC).

Приложения като автономна мобилност, индустриална автоматизация, виртуална и добавена реалност (VR/AR) и холографско телеприсъствие налагат изключително строги изисквания към качеството на услугата (QoS) и качеството на потребителското изживяване (QoE). Традиционните мрежови архитектури за радиодостъп (RAN) са монолитни, зависими от специализиран хардуер и със затворени интерфейси, което ограничава тяхната гъвкавост и забавя иновациите.

В отговор на тези ограничения се появява концепцията за мрежа за достъп с отворени интерфейси (Open RAN), която въвежда принципите на дезагрегация и виртуализация. Динамичната природа на радиосредата и хетерогенността на услугите обаче правят традиционните методи за управление неефективни, което налага интегрирането на изкуствен интелект (AI) и машинно обучение (ML) за постигане на автоматизация в реално време и динамично разпределение на мрежовите ресурси.

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на дипломната работа е да предложи цялостна методология за подобряване на QoS и QoE, чрез интегриране на изкуствен интелект (AI) в мрежовите операции за оптимизиране на тяхната производителност. Мрежовите операции, базирани на изкуствен интелект, са интегрирани в Open RAN архитектурата, за да поддържат различни случаи на употреба с хетерогенни изисквания за QoS/QoE по отношение на честотна лента, латентност, загуба на пакети и джитер. Разработени са алгоритми за машинно обучение (ML), за да осигурят автоматизирано прогнозиране на мрежовия трафик и оптимално разпределение на ресурсите чрез използване на техники за нарязване на мрежата (Network Slicing). Прилагането на автоматизация на оркестрация на ресурсите, подпомагана от ML, води до оптимално управление на мрежата и предоставя нова и ефективна методология за намаляване на капиталовите разходи (CAPEX) и оперативните разходи (OPEX) на доставчиците на комуникационни услуги (CSP). За постигане на поставената цел са дефинирани следните задачи:

1. Анализ и систематизация на архитектурната еволюция на RAN и прехода към отворени и виртуализирани архитектури (C-RAN, vRAN, O-RAN), както и ролята на интелигентните контролери (RIC).
2. Проектиране и реализация на пълнофункционална експериментална O-RAN тестова среда, базирана на софтуер с отворен код и COTS хардуер.
3. Разработване на алгоритми за откриване на аномалии чрез модели за дълбоко обучение за проактивно идентифициране на нетипично поведение в мрежовия трафик.

4. Моделиране и прогнозиране на QoS и QoE при интерактивни мултимедийни услуги (гейминг, VR) и C-V2X комуникации чрез машинно обучение.
5. Оптимизиране на мрежовите ресурси чрез механизми за интелигентна класификация на трафика и динамично разпределение на радиоресурсите.

Научна новост

Научната новост на дисертационния труд се състои в:

- Предлагане на методология за управление на мрежата, специално насочена към QoS-критични приложения, чрез интегриране на AI алгоритми в интелигентния контролер на радиомрежата (RIC).
- Разработване и внедряване на иновативен Трансформаторен модел (Transformer) за откриване на сложни мрежови аномалии, който превъзхожда традиционните методи и LSTM архитектурите по точност и време за обучение.
- Създаване на специализирани прогностични модели за QoE и QoS:
    - Multi-headed CNN архитектура за предсказване на качеството при гейминг видео стрийминг
    - LSTM encoder-decoder модел за VR 360-градусово видео, улавящ дългосрочни времеви зависимости.
    - Локационно-независим (location-agnostic) подход за прогнозиране на QoS в C-V2X сценарии, позволяващ висока степен на генерализация между различни мобилни оператори
- Успешна имлементация на O-RAN базирана методология за адаптивно и динамично разпределение на физически ресурсни блокове между мрежови парчета, която осигурява гарантирано качество за UHD видео потоци.

Практическа приложимост

Всички разработени методи и алгоритми, както и предложените подобрения към вече практически имплементирани такива, са изследвани и анализирани посредством симулационни експерименти. Направено е и сравнение с други съществуващи модели, в основата на които са заложени подобни функционалности и характеристики, или имат сходни цели по отношение на подобрение на работните им параметри. Всичко това прави възможността за внедряване на резултатите от настоящия дисертационен труд непосредствена и лесно реализуема в съвременните телекомуникационни мрежи. Допълнително доказателство за тази висока приложимост е успешното интегриране на предложените алгоритми за машинно обучение под формата на микроуслуги в реални мрежови компоненти, като функцията за анализ на мрежови данни и интелигентните контролери на радиомрежата, което категорично потвърждава тяхната оперативна съвместимост и готовност за практическо приложение .

Публикуване на резултатите от дисертационното изследване

Направените анализи, предложените подходи и получените резултати за периода 2021÷2025 са представени в общо **14** авторски публикации индексирани в Scopus и Web of Science: **1** глава от книга; **9** публикации в *международни конференции*; **4** публикации в *международни*

научни списания с ранг Q1 и Q2. Статиите имат общо 59 цитирания в SCOPUS и 64 цитирания в Google scholar.

Структура и обем на дисертационния труд

Дисертационният труд е написан на български език и е в обем от **193** страници формат А4 и съдържа увод, осем глави, заключение с изложени основни приноси, списък на фигурите, списък на таблиците, списък на използваните съкращения, списък с публикациите по дисертацията, списък на използваната литература и две приложения. Изложението на дисертационния труд съдържа **70** фигури и **13** таблици. Използвани са **138** литературни източника като всички са на латиница и над 80% са от последните десет години. Номерата на фигурите и таблиците в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

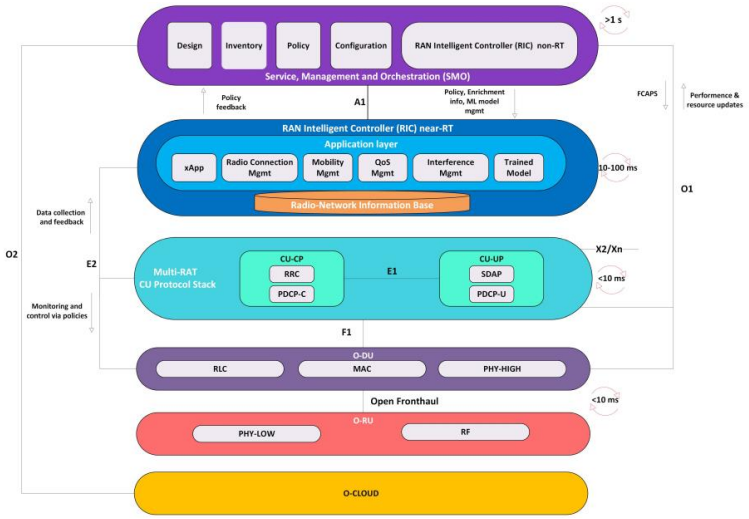
1. Еволюция на мрежите за радиодостъп и техните архитектури

Първата глава на дисертационния труд представя изчерпателен анализ на историческото развитие и архитектурната еволюция на мрежите за радиодостъп, като се проследява пътят от ранните цифрови системи до съвременните отворени и дезагрегирани среди. Изложението започва с разглеждане на архитектурите от второ поколение, като се обяснява структурата на подсистемата на базовите станции в GSM и GPRS[1]. Посочва се, че въвеждането на пакетната комутация чрез GERAN [2] бележи важен етап, но архитектурата остава ограничена поради своята монолитност и зависимост от специфичен хардуер. При анализа на третото поколение се разглежда ролята на UTRAN [4], където управлението на радиоресурсите става по-комплексно поради използването на WCDMA, изискващо нови механизми за контрол на мощността и мобилността.

Особено внимание в главата е отделено на прехода към четвърто поколение LTE и неговата мрежа за достъп E-UTRAN[5]. Тук детайлно се изследва концепцията за плоска IP архитектура, при която функциите на контролния възел са интегрирани директно в базовата станция eNodeB. Този подход значително намалява латентността и опростява мрежовата топология, което е критична стъпка към поддръжката на съвременните мобилни услуги. В дисертацията се проследява как тази децентрализация подготвя почвата за пето поколение мобилни мрежи, където базовата станция gNodeB[8] вече се разглежда като съвкупност от логически дезагрегирани единици. Разглежда се функционалното разделяне на централни и разпределени блокове, което позволява изключителна гъвкавост при разполагането на мрежовите функции в зависимост от конкретните изисквания на приложенията за капацитет и закъснение.

Важна част от изложението заема и анализът на парадигмите Cloud RAN [17] и Software-Defined RAN [21]. Изследват се предимствата на централизираната обработка на сигналите в BBU клъстерите (BBU pools), което позволява по-добра координация на ресурсите и намаляване на разходите. Същевременно са отбелязани и ограниченията на тези системи, свързани с високите изисквания към капацитета на транспортната мрежа и използването на затворени интерфейси. Това логически обосновава необходимостта от Open RAN архитектура, която се дефинира като еволюционна стъпка към пълна оперативна съвместимост между различни производители.

Главата завършва с подробен технически обзор на архитектурата на O-RAN [30]. Дефинирани са и основните компоненти: Near-Real-Time RIC, Non-Real-Time RIC, O-CU и O-DU, като се обяснява тяхното взаимодействие чрез отворените интерфейси A1, E2 и O1. Подчертава се, че интелигентността е вградена в самата структура на мрежата чрез тези контролери, което позволява внедряването на xApps и rApps за динамична оптимизация. Този детайлен архитектурен разбор служи за основа на следващите глави, в които се предлагат конкретни алгоритми за машинно обучение, интегрирани в описаната рамка. В заключение на първа глава се прави изводът, че отворената и дезагрегирана архитектура е задължително условие за реализация на услуги с критични изисквания към качеството на обслужване в хетерогенната среда на бъдещите мрежи.

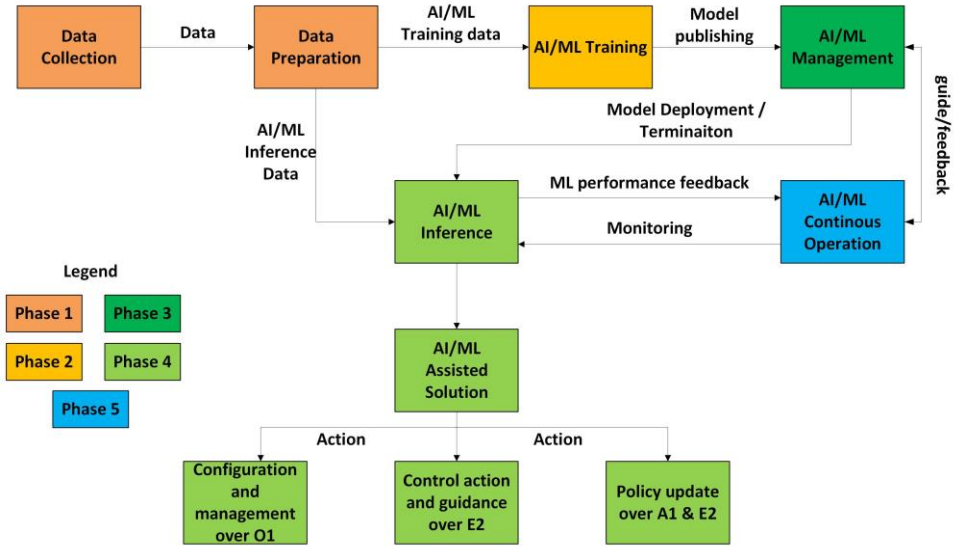


Фигура 1.4. Архитектура на O-RAN

2. Приложение на машинното обучение в клетъчните мрежи

Втората глава на дисертационния труд е посветена на задълбочено изследване на теоретичните основи и практическите аспекти на интегрирането на изкуствения интелект и машинното обучение в съвременните мрежи за радиодостъп. Изложението започва с детайлна таксономия на алгоритмите за машинно обучение, като се прави критичен преглед на тяхната приложимост за оптимизация на мрежовите операции. Разгледани са основните парадигми на обучението с надзор (Supervised learning) [35], обучението без надзор (Unsupervised learning) [36] и обучението чрез стимулиране (Reinforcement learning) [37], като за всяка от тях са посочени конкретни сценарии на употреба в клетъчните системи. Подчертава се, че докато обучението с надзор е незаменимо за задачи по класификация и регресия при наличие на етикетирани исторически данни, то обученията без надзор и чрез стимулиране предоставят уникални възможности за вземане на решения в реално време при динамично променящи се радиоусловия.

Важен акцент в главата е поставен върху жизнения цикъл на моделите за машинно обучение в рамките на O-RAN архитектурата, съгласно спецификациите на O-RAN Алианса [43]. Детайлно са описани процесите на събиране на данни, предварителна обработка, обучение, разгръщане и мониторинг на моделите. Разглежда се ролята на Non-Real-Time RIC за офлайн обучение и управление на политиките, както и на Near-Real-Time RIC за изпълнение на обучените модели под формата на xApps за контрол в рамките на милисекунди. Обсъждат се и предизвикателствата, свързани с оперативната съвместимост на модели от различни производители и необходимостта от стандартизирани интерфейси за обмен на данни (като E2 и A1), за да се гарантира затворен цикъл на управление и самооптимизация на мрежата.



Фигура 2.2. Жизнен цикъл на ML/AI модела в O-RAN

Главата завършва с преглед на методите за оценка на точността и надеждността на предложените модели. Авторът анализира различни метрики за производителност и подчертава значението на генерализацията на моделите, за да се осигури тяхната работоспособност в различни географски региони и при разнообразни конфигурации на базовите станции. Формулирани са изводи за необходимостта от хибридни подходи, съчетаващи експертни знания за физическия слой с гъвкавостта на дълбокото обучение. Този теоретичен фундамент служи за логическа връзка към следващите глави, където разработените модели се прилагат за решаване на конкретни проблеми, свързани с откриване на аномалии и прогнозиране на качеството на обслужване.

3. Дефиниране на услуги, критични към качеството на обслужване

Третата глава на дисертационния труд е посветена на детайлното изследване и дефиниране на съвременните мултимедийни и интерактивни услуги, които поставят най-високи изисквания към производителността на безжичните мрежи. Изложението започва с преглед на фундаменталната промяна в парадигмата за оценка на качеството, като обосновава прехода от чисто технически показатели (QoS) към комплексното качество на

потребителското изживяване (QoE). Посочва се, че традиционните показатели за качество на услугата, макар и необходими за мониторинг на мрежата, често не успяват да отразят реалното удовлетворение на крайния потребител, особено при силно динамични услуги като виртуалната реалност и облачен гейминг. В този контекст се анализират субективните фактори и психологическите аспекти на възприятието, които определят крайните стойности на средното мнение [52].

Специално внимание е отделено на спецификата на виртуалната реалност и нейните изисквания за ултраниска латентност и висока честотна лента. Изследва се концепцията за потапяне и присъствие във виртуалната среда, като дефинира критичните прагове за закъснение, над които се появява ефектът на симулаторна болест (cybersickness). Анализирани са различните видове закъснения в системата, включително времето от движение до фотон, и как те влияят на общото качество на изживяването при 360-градусов видео стрийминг. В дисертацията се предлага класификация на факторите на влияние, разделени на системни, човешки и контекстуални, като се проследява тяхното индивидуално и комбинирано въздействие върху крайния потребител [57].

Втората голяма група услуги, разгледани в главата, са онлайн игрите и облачният гейминг. Подчертава се, че за разлика от пасивното видео съдържание, интерактивността тук въвежда нови зависимости, при които дори малки флуктуации в мрежовите параметри (джитер) могат драстично да влошат игровия процес [61]. Изследвана е връзката между сложността на видео съдържанието, динамиката на сцените и необходимия битрейт за поддържане на задоволително качество при високи резолюции. Изложението обхваща и сценариите за комуникация между превозни средства (C-V2X), където надеждността и скоростта на предаване имат критично значение за безопасността на движението, което налага строги изисквания към достъпността и капацитета на радиоканала [71].

Главата завършва с анализ на математическите модели за картографиране между показателите за качество на услугата и качеството на изживяването. Разглеждат се и функциите за нелинейна зависимост, като логаритмичния закон на Вебер-Фехнер и експоненциалната хипотеза IQX, които служат за основа на разработените в следващите глави прогностични модели. Направени са изводи за необходимостта от адаптивни механизми за управление, които да отчитат специфичния профил на всяка услуга в реално време. Този анализ служи за дефиниране на входните параметри и целите на оптимизацията, които се използват при проектирането на интелигентната O-RAN среда и последващите експерименти за подобряване на качеството на обслужване [56].

4. Проектиране и имплементиране на 5G/LTE тестова мрежа базирана на O-RAN стандарта

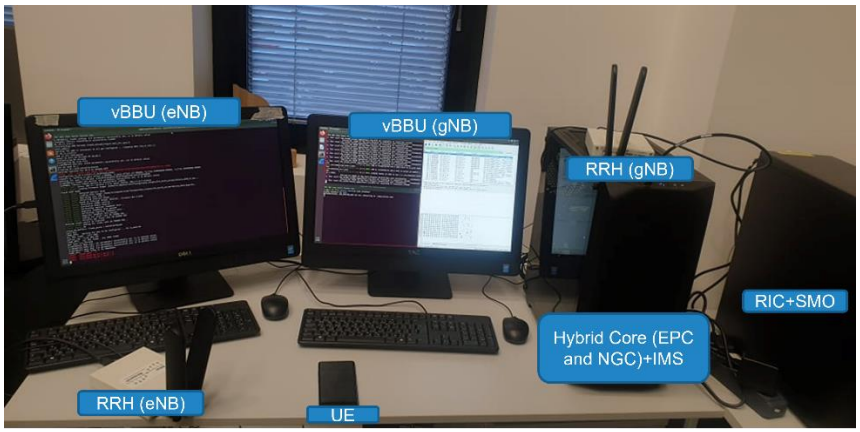
Разбирането на архитектурните концепции, приложени във водещи широкомащабни 5G/LTE тестови платформи, предоставя ценна теоретична и технологична рамка за проектирането на собствени тестови мрежи. В четвърта глава на дисертацията са анализирани две от най-широко използваните тестови платформи в световен мащаб: COLOSSEUM [76] и POWDER [92]. Те служат като два забележителни примера, които нагледно демонстрират как фундаменталните принципи на хардуера с общо предназначение (COTS), софтуеризацията, отворения код, виртуализацията и контейнеризацията са успешно интегрирани за постигане на реалистични, мащабируеми и напълно програмируеми изследователски среди. COLOSSEUM, като най-големият в света безжичен мрежов емулатор, действащ като високоточен дигитален близнак на O-RAN, предоставяйки възможност за безопасно генериране на данни и обучение на AI/ML модели чрез сложна система за канална емулация и мащабна изчислителна инфраструктура. От друга страна, платформата POWDER представлява мащабна "жива лаборатория" в градска среда, която балансира между суров достъп до хардуера за фундаментални изследвания и абстракция за по-високите мрежови слоеве чрез богата екосистема от софтуер с отворен код и гъвкава оптична транспортна мрежа. Именно въз основа на този задълбочен анализ и извлечените от него най-добри архитектурни практики е проектирана, аргументирана и реализирана собствената експериментална O-RAN тестова мрежа, която да служи като фундамент за валидация на предложените в дисертацията алгоритми.

При проектирането й са спазени основните концепции и архитектурни принципи за отвореност и независимост от конкретен производител, като мрежата е изградена изцяло с използване на софтуер с отворен код и COTS хардуер. Мрежата поддържа работа с всички стандартизирани честоти от Frequency Range 1 (FR1), като експериментите са фокусирани основно върху Band 7 (2.6 GHz) за LTE и Band n78 (3.5 GHz) за 5G NR. Разработената мрежа предоставя широк спектър от телекомуникационни услуги, демонстрирайки възможностите

на хибридна 4G/5G среда. Поддържат се услуги за мобилен широколентов достъп с пакетна комутация и QoS диференциация, изпращане на кратки съобщения чрез SGs интерфейс към 3G компоненти на опорната мрежа, както и богат набор от IMS-базирани услуги чрез интегрирана open-source платформа, включително VoLTE и SMS over SIP. По отношение на 5G, мрежата поддържа високоскоростен достъп (eMBB) както в преходния неавтономен режим (Non-standalone; NSA), използващ 4G опорна инфраструктура, така и в пълноценен автономен режим (Standalone; SA) с независима 5G опорна мрежа.

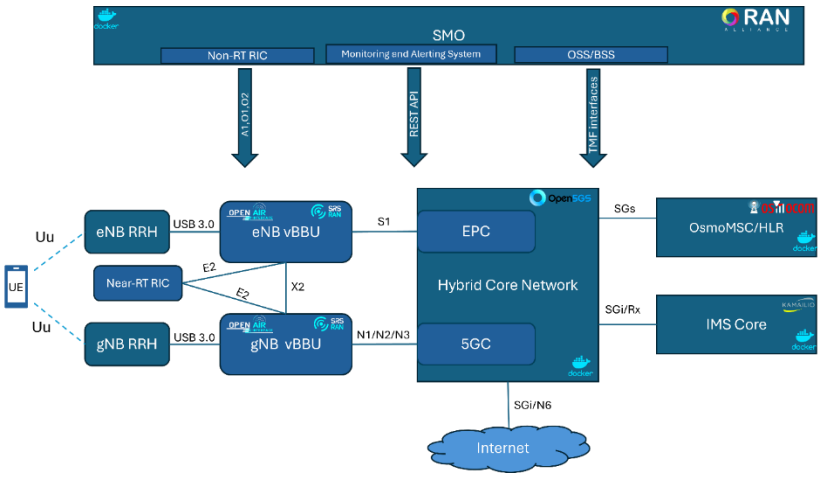
a. Архитектура на мрежата

Физическата архитектура на тестовата мрежа е изградена около стандартен настолен компютър, оборудван с процесор Intel Core i9 от десето поколение, 64 GB оперативна памет, графичен процесор NVIDIA GTX 1650 и операционна система Ubuntu. Този възел хоства всички основни компоненти на мрежата, включително елементите на мрежата за радиодостъп за 5G gNB и LTE eNB, хибридната опорна мрежа и IMS платформата. За реализация на интелигентното управление е използван отделен сървър, на който са конфигурирани две виртуални машини чрез хипервайзор VMware vSphere. Първата виртуална машина хоства компонентите на контролера работещ в близко до реалното време (Near-RT RIC), докато втората е предназначена за функционалността за управление и оркестрация на услугите (SMO). Радио интерфейсът е реализиран чрез две софтуерно дефинирани радиа USRP B210, свързани към хост системата чрез USB 3.0 интерфейс, които осигуряват 2x2 MIMO конфигурация и поддържат широк честотен обхват. За валидиране на функционалността са използвани разнообразни потребителски устройства, включително смартфони, LTE USB донгъли и специализиран 5G модул Quectel RM500Q-GL, предоставящ достъп до детайлна телеметрия.



Фигура 4. 1. Физическа архитектура на тестовата мрежа

Логическата архитектура представлява комплексна имплементация на O-RAN, интегрираща множество софтуерни компоненти в съответствие със спецификациите на O-RAN Алианса и 3GPP. Мрежата за радиодостъп поддържа едновременна работа на LTE и 5G технологии, като базовите станции могат да функционират както в традиционен монолитен режим, така и с функционално разделение по Опция 2 (при границата PDCP/RLC) и Опция 7.2 (разделяне между нисък и висок физически слой). RAN мрежата е реализирана чрез две алтернативни софтуерни платформи: srsRAN, която функционира в контейнеризирана среда чрез Docker, и OpenAirInterface (OAI), инсталирана директно върху операционната система на хост машината (bare metal конфигурация), което позволява оптимална производителност и директен достъп до хардуерните ресурси. Опорната мрежа е базирана на Open5GS, интегрираща EPC и 5GC функционалности, и е допълнена с Kamailio IMS платформа и Osmocom компоненти за реализиране на допълнителни телеком услуги. Всички тези компоненти на опорната мрежа са контейнеризирани и обединени в обща Docker Compose среда. Интелигентното управление в близко до реалното време (Near-RT RIC) е реализирано чрез три алтернативни имплементации: FlexRAN (в bare metal конфигурация), FlexRIC (контейнеризиран чрез Docker) и официалната референтна имплементация OSC Near-RT RIC, реализирана като микроуслуги върху Kubernetes клъстер. Цялостната оркестрация е осигурена от OSC SMO, който също е базиран на микроуслуги и Kubernetes, интегрирайки Non-RT RIC компонентите. Взаимодействието между всички тези елементи се осъществява чрез стандартизираните интерфейси E2, X2, S1, N1/N2/N3 и O1.



Фигура 4. 2. Логическа архитектура на тестовата мрежа

b. Оценка на работоспособността на мрежата

Представените експерименти са проведени без интегрирани компоненти за интелигентно управление, като Near-RT RIC и SMO, поради липса на релевантност към базовите тестове. За оценката са избрани ключови показатели за ефективност, които са от първостепенна важност за приложения, критични към качеството на услугата. Тези показатели включват времето за двупосочно предаване (Round Trip Time, RTT), измерено чрез изпращането на ICMP пакети до сървър в интернет (от край до край) и в рамките на радиомрежата (RAN), пропускателната способност в права и обратна посока чрез инструменти за тест на скоростта, както и колебанията в закъснението на пакетите (джитер) при UDP свързаност.

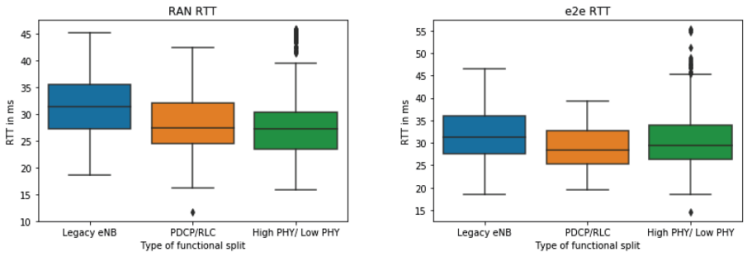
i. Оценка на работоспособността на LTE мрежата

Експериментите за установяване на базовите показатели за производителност са проведени върху LTE конфигурацията на тестовата мрежа, като за потребителско устройство е използван смартфон Samsung Galaxy Note 9. Измерванията обхващат три архитектури на базовата станция: традиционна монолитна eNB архитектура без функционално разделяне, функционално разделяне по Опция 2 (на границата PDCP/RLC) и функционално разделяне по Опция 7.2 (разделяне между нисък и висок физически слой). Параметрите на мрежата включват честотна лента от 20 MHz, съответстваща на 100 ресурсни блока в режим на честотно разделяне (FDD). Работните честоти са 2.67 GHz в права посока с 64QAM модулация и 2.56 GHz в обратна посока с QPSK модулация при един свързан потребител.

Параметри	Стойности
Честотна лента	20 MHz (100 ресурсни блока)
Честота в права посока	2.67 GHz
Честота в обратна посока	2.56 GHz
Техника за разделяне на спектъра	FDD
Модулация в права посока	64QAM
Модулация в обратна посока	QPSK
Режим на предаване	1
Брой свързани потребители	1

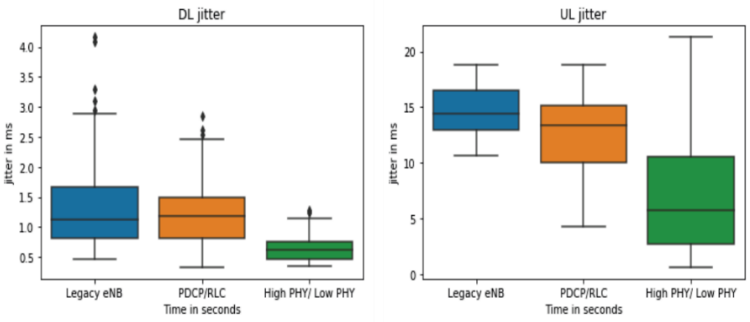
Таблица 4. 1. Параметри на LTE мрежата

При цялостната оценка на времето за двупосочно предаване (RTT) са изпратени 1000 ICMP пакета за измерване на латентността от край до край и в рамките на RAN мрежата. Резултатите разкриват, че традиционната монолитна eNB архитектура последователно показва повишени нива на латентност, надвишавайки стойностите на дезагрегираните Опция 2 и Опция 7.2 с 15-20%. Двете опции за функционално разделяне демонстрират сходни базови характеристики с разлики в RTT, ограничени до 1-2 ms. Въпреки това, при Опция 7.2 се регистрират съответно 1.8% и 5.4% пакети с абнормални RTT стойности при измерванията от край до край и в RAN сегмента. Тази честота на отклоненията прави Опция 2 по-предпочитана за приложения с твърди изисквания в реално време, тъй като предоставя по-надеждна производителност в 99-ия персентил.



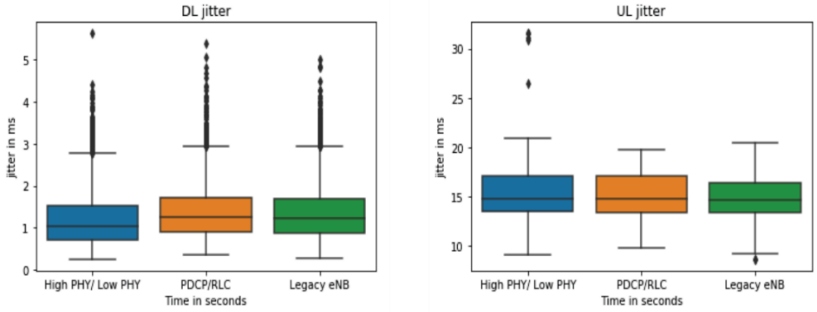
Фигура 4. 3. RTT при различните видове eNB архитектури

Оценката на джитера, осъществена чрез 60-секундна UDP връзка, показва, че Опция 7.2 постига най-ниските стойности в права посока, достигайки под 1 ms. Наблюдава се зависимост, според която централизирането на повече мрежови функции (както е при Опция 7.2) води до по-тесен диапазон на вариациите в латентността и по-стабилна производителност.



Фигура 4. 4. Jitter в права (DL) и обратна (UL) посока при 60 секундна връзка

За потвърждаване на тази тенденция са проведени и дългосрочни 60-минутни измервания. При тях първоначалното предимство на Опция 7.2 се запазва леко, но разликите между трите архитектури стават статистически незначителни във времето. Това показва, че и трите архитектурни подхода осигуряват приемливи стойности на джитер за продължително функциониране на чувствителните към джитер приложения.



Фигура 4. 5. Jitter в права (DL) и обратна (UL) посока при 60 минутна връзка

Анализът на пропускателната способност демонстрира, че медианните скорости в права посока (DL) са сравними при всички конфигурации. За разлика от това, в обратна посока (UL) се наблюдава драматично подобрение от над 4 пъти при функционалните разделения спрямо традиционната монолитна архитектура. Това може да се обясни с оптимизираното управление на ресурсите и по-ефективното планиране в централизираните модули на сплит архитектурите. Комбинацията от значително повишена UL скорост и по-ниска латентност прави дезагрегираните решения изключително подходящи за съвременни симетрични услуги като холографски видеоконференции и интерактивни системи, изискващи оптимална и предсказуема комуникация.

	Legacy eNB				PDCP/RLC				High PHY/ Low PHY			
No.	Ping (ms)	Jitter (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	Jitter (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	Jitter (ms)	DL (Mbps)	UL (Mbps)
1	27	3	65.9	4.8	28	18	64.2	19.8	27	2	69.6	18.33
2	12	20	16.1	4.02	18	17	66.2	19.1	27	2	66.2	17.9
3	21	5	63	4.67	27	3	61.9	18.3	27	9	68.3	18.3
4	28	19	68.4	4.44	28	1	66.5	18.3	21	9	66.8	18.33
5	22	10	61.2	4.35	26	4	66.8	9.87	28	1	68.5	18.2
6	27	23	32.3	4.23	28	1	67.6	18.3	2	2	16.08	18.3
Средно	22.83	13.33	51.15	4.42	25.83	7.33	65.53	17.28	26.33	4.17	59.25	18.23
Медиана	24.5	14.5	62.1	4.395	27.5	3.5	66.35	18.3	27	2	67.55	18.3

Таблица 4. 2. Пропускателна способност на мрежата в LTE режим

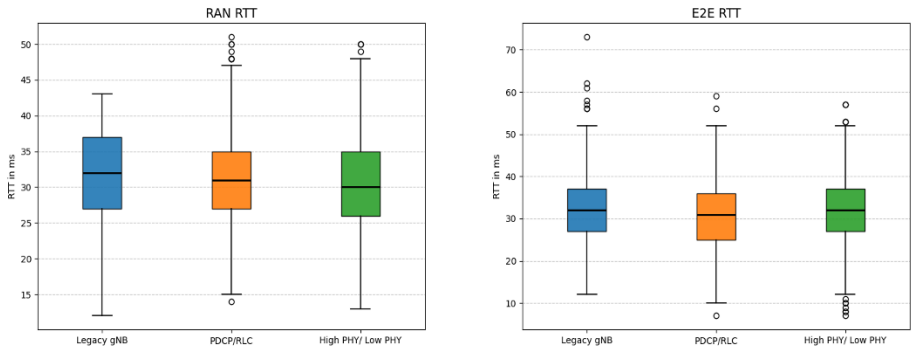
c. Оценка на работоспособността на 5G мрежата в SA режим

Експериментите са проведени и с тестовата мрежа в конфигурация на 5G SA мрежа, за да се оцени влиянието на различните архитектурни варианти на базовата станция (gNB) върху ключовите показатели за производителност. Използвани са същите три архитектури като при LTE тестовете: традиционна монолитна, Опция 2 и Опция 7.2. За да се осигури директна съпоставимост, мрежовите параметри са идентични за всички конфигурации и са базирани на спецификациите на 3GPP. Конфигурацията включва 40 MHz честотна лента с честота 3.5 GHz в режим на дуплекс с времево разделяне (TDD), използвайки 64QAM модулация в права посока и QPSK в обратна посока при един свързан потребител.

Параметри	Стойности
Честотна лента	40 MHz (106 ресурсни блока и 30 KHz SCS)
Честота в права посока	3.5 GHz
Честота в обратна посока	3.5 GHz
Техника за разделяне на спектъра	TDD
Модулация в права посока	64QAM
Модулация в обратна посока	QPSK
Режим на предаване	1
Брой свързани потребители	1

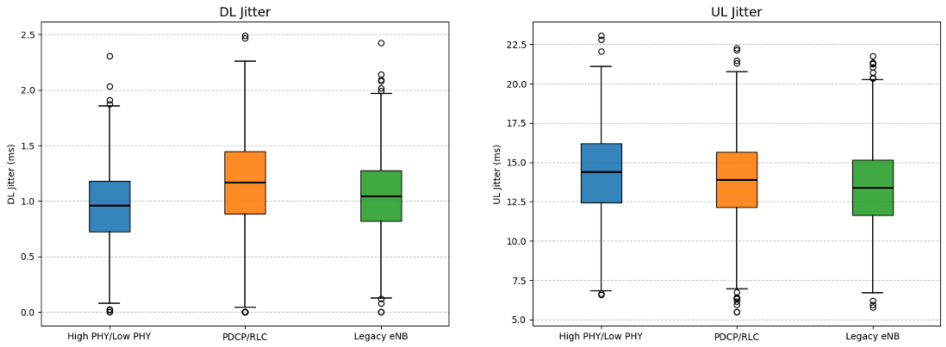
Таблица 4. 3. Параметри на 5G мрежата

Оценката на латентността отново е осъществена чрез измервания на времето за двупосочно предаване (RTT) на две нива: в рамките на радиомрежата (RAN RTT) и от край до край до външен интернет сървър (e2e RTT). Получените резултати показват леко предимство на функционалните разделения спрямо традиционната монолитна архитектура. При RAN RTT двете сплит конфигурации демонстрират по-ниски стойности, като при Опция 7.2 намаляването достига до приблизително 6%, докато разликите между Опция 2 и Опция 7.2 остават почти незначителни. При измерванията от край до край Опция 2 показва леко предимство, докато Опция 7.2 достига медианни стойности, близки до традиционната архитектура, но с малко по-високи отклонения в горния диапазон. Тези вариации могат да се обяснят със специфичните механизми на разделението във физическия слой или на ограничения в производителността на използвания хардуер.



Фигура 4. 6. RTT при различните видове gNB архитектури

Измерванията на джитера са проведени чрез дългосрочни 60-минутни UDP сесии за двете посоки на предаване. В права посока (DL) Опция 7.2 демонстрира най-ниски стойности на джитер, които в някои случаи са близо 1.5 пъти по-ниски спрямо монолитната архитектура. Опция 2 също отчита умерено подобрение, което доказва, че централизацията на по-голям обем функционалности и постоянният битрейт във фронтхола водят до по-стабилна латентност в низходяща посока. В обратна посока (UL) обаче, Опция 7.2 регистрира чувствително по-високи стойности на джитер спрямо останалите архитектури. Това потвърждава, че джитерът е особено чувствителен към разпределението на функциите на физическия слой и че по-високата степен на централизация изисква правилно функциониране за постигане на стабилност.



Фигура 4. 7. Стойности на Jitter в права и обратна посока при различните gNB архитектури

По отношение на измерената пропускателна способност, разликите между трите архитектури се оказват незначителни, като всички осигуряват стабилна производителност, която напълно очаквано превъзхожда скоростите, постигнати в LTE режим.

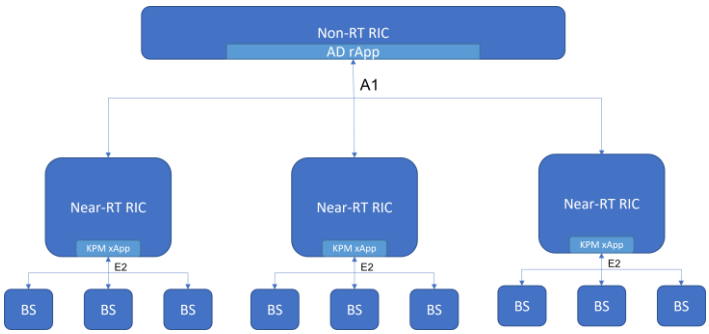
	Legacy gNB			PDCP/RLC			High PHY/ Low PHY		
No	Ping (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	DL (Mbps)	UL (Mbps)
1	20	81.95	36.65	20	86.05	38.48	24	84.33	37.71
2	20	77.22	36.59	20	81.13	38.42	22	81.27	37.57
3	20	85.03	36.53	22	86.73	37.27	21	83	36.52
4	20	80.34	36.13	20	78.73	36.5	23	77.94	36.87
5	20	82.01	36.53	24	81.97	37.27	20	82.77	37.64
6	20	83.61	36.43	20	84.28	37.16	20	81.13	36.79
Avg	20	81.69333	36.47667	21	83.14833	37.51667	21.66667	81.74	37.18333
Median	20	81.98	36.53	20	83.125	37.27	21.5	82.02	37.22

5. Автоматизирано откриване на аномалии в мобилните мрежи

Пета глава на дисертационния труд е посветена на проектирането, внедряването и оценката на интелигентни системи за откриване на аномалии, които са критични за осигуряването на сигурността и надеждността на мобилните мрежи. Теоретичното въведение в главата обосновава необходимостта от автоматизирани подходи поради нарастващата сложност на мрежовите архитектури и невъзможността за ръчен мониторинг на огромните обеми от генерирани данни. Разгледана е таксономията на аномалиите, включваща точкови, контекстуални и колективни аномалии, като се подчертава, че в съвременните телекомуникационни системи те често са индикатори за кибератаки, хардуерни откази или софтуерни грешки. Преходът от традиционни статистически методи към модели, базирани на дълбоко обучение, се дефинира като задължително условие за ефективно разпознаване на сложни и непознати досега модели на нетипично поведение в мрежовия трафик. Главата също така представя работата от три научни публикации, които обхващат различни аспекти на проблема от теоретичния анализ и сравнение на подходи до тяхната практическа имплементация.

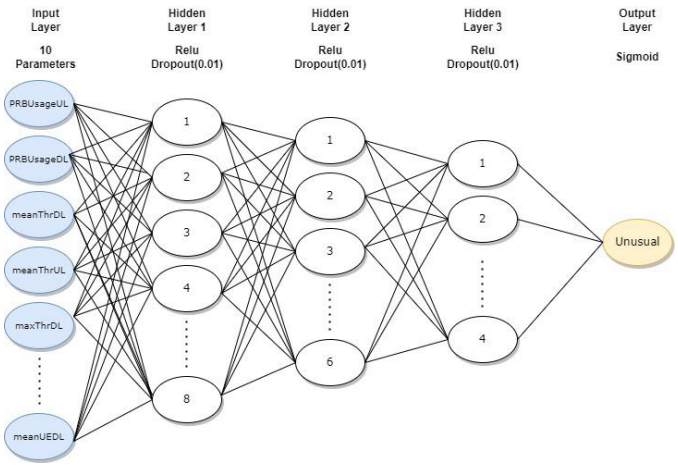
а. Разработване и имплементация на алгоритми за навременно откриване на аномалии в мобилни мрежи

Първо е представен алгоритъм за дълбоко обучение за откриване на аномалии, който може да се изпълнява като rApp приложение върху компонента Non-RT RIC на архитектурата O-RAN. Използвайки специфичните за мрежата за радиодостъп ключови показатели за ефективност (KPI), получени от E2 възлите, rApp приложението следи дългосрочните тенденции и модели по отношение на производителността и обучава модел за дълбоко обучение с надзор въз основа на тях. При откриване на необичайно поведение на мрежата или неоптимална производителност, приложението може да предприеме коригиращи действия чрез изпращане на инструкции за преконфигуриране през A1 интерфейса до Near-RT RIC, който от своя страна контролира базовата станция, където е възникнала аномалията.



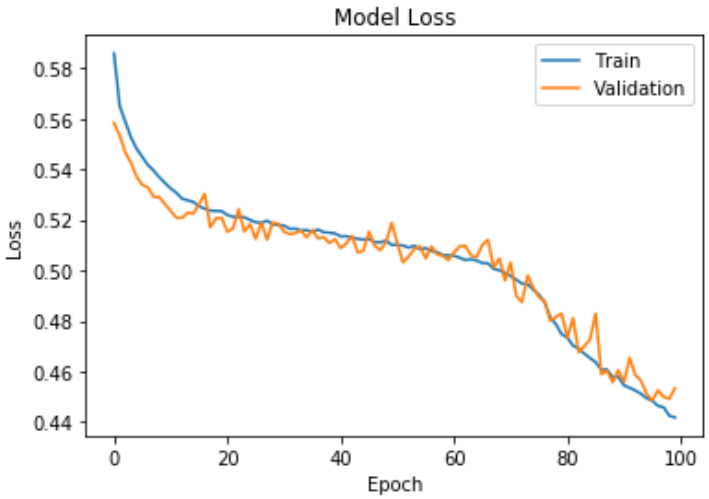
Фигура 5. 1. Схема на rApp за откриване на аномалии

Разработен е прототип на предложеното решение, използвайки JupyerNotebook, Python 3.7.2, Keras и Tensorflow. Оценката на производителността е проведена в типичен сценарий, включващ мониторинг на мрежата чрез KPM xApp, работещ върху Near-RT RIC. Данните, използвани за обучението на модела, са събрани от реална LTE мрежа и се състоят от двуседмични записи от група от общо 10 базови станции с интервал на отчитане от 15 минути. Въпреки че данните произхождат от традиционна LTE архитектура, същите параметри могат да бъдат записани по време на цикъла за мониторинг в O-RAN мрежи. Всяка проба в набора от данни съдържа характеристики като времева марка, уникален идентификатор на клетката, процентно използване на физическите ресурсни блокове (PRB) в права и обратна посока, среден и максимален пренесен трафик (в Mbps), както и среден и максимален брой едновременно активни потребителски устройства. За целите на обучението с надзор е включен и етикет, при който стойност нула указва нормална работа, а единица указва необичайно поведение. Окончателната архитектура на невронната мрежа е избрана след основно фино настройване на хиперпараметрите.



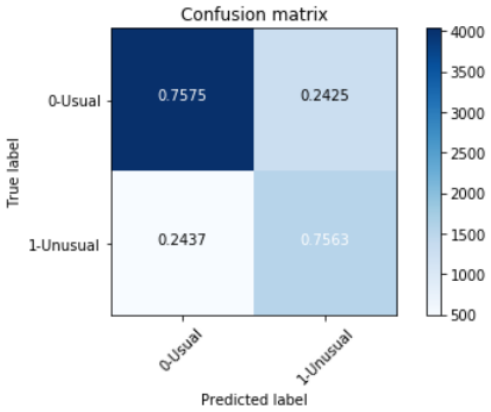
Фигура 5. 2. Архитектура на избраната невронна мрежа

Архитектурата на крайния модел се състои от входен слой, който приема и обработва суровите данни (с изключение на времевата марка, името на клетката и максималния брой активни устройства общо за двете посоки), последван от три скрити слоя. Тези слоеве съдържат съответно 8, 6 и 4 неврона, оборудвани с изчислително ефективната активираща функция Rectified Linear Unit (ReLU), която помага за улавянето на сложни модели и смекчава проблема с изчезващия градиент. За предотвратяване на прекомерното приспособяване (overfitting), след всеки скрит слой са добавени слоеве за отпадане (Dropout), които произволно деактивират част от невроните по време на обучението, подобрявайки генерализацията на модела. Изходният слой се състои от един неврон със сигмоидна активационна функция, която преобразува изхода във вероятностна оценка за бинарна класификация, показвайки вероятността входната проба да представлява аномалия.



Фигура 5. 3. Загуби на модела при процеса на обучение и валидация

Преди обучението наборът от данни е разделен на 70% за обучение, 10% за кръстосана валидация и 20% за тестване. Моделът е обучаван в продължение на 100 епохи с размер на партидата от 32 примера и темп на обучение 0.01. Анализът на графиката на загубите показва, че загубите от обучението намаляват с всяка епоха, което потвърждава ефективното учене. Загубата при валидиране първоначално следва подобна тенденция, като се наблюдават незначителни отклонения от около 1%, които не влияят съществено на модела.



Фигура 5. 4. Матрица на грешките (Confusion Matrix)

След оценка на ефективността, алгоритъмът постига обща точност от 0.7571, прецизност от 54.43%, чувствителност (Recall) от 0.7563 и F1 резултат от 0.6630. Въпреки че матрицата на грешките разкрива добре балансирано съотношение между фалшивите положителни и фалшивите отрицателни резултати, моделът се оказва неефективен за практически цели поради големия брой грешни класификации. Ниската прецизност показва, че голям обем нормален трафик би бил класифициран като анормален, което би претоварило мрежата с ненужни съобщения за реконфигурация, докато фалшиво отрицателните резултати биха довели до недиагностицирана деградация на критични услуги.

Друг съществен недостатък на предложения подход е неговата зависимост от обучение с надзор. За да функционира ефективно, този метод изисква предварително етикетирани набори от данни, които са изключително трудни за събиране в динамични мрежови среди поради присъщия класов дисбаланс, при който нормалните данни винаги доминират над редките аномалии. Това ограничение налага преминаването към подходи, базирани на обучение с частичен надзор или без надзор. Тези методи се обучават само върху нормални оперативни данни и са способни да откриват както познати, така и нови, нетипични аномалии. Следователно, бъдещите изследвания са фокусирани върху разработката и прилагането на алгоритми за обучение без надзор, които осигуряват по-висока степен на автоматизация и надеждност при управлението на съвременните мобилни мрежи.

b. Модел за откриване на аномалии, базиран на реконструкция чрез Трансформатор

Във втората част на главата е представен модел, базиран на реконструкция, който използва архитектурата на Трансформатор (Transformer) за откриване на аномалии в оперативните данни на мобилна мрежа. Разработката на модела е мотивирана от доказания успех на Трансформаторите при откриването на аномалии в ЕКГ данни [124], където тяхната способност да улавят необичайни модели както на локално, така и на глобално ниво, се оказва изключително ефективна. За сравнение на резултатите от трансформатора е използвана стандартната архитектура LSTM-Autoencoder, която следва същия принцип на реконструкция. Архитектурата се състои от енкодер с два LSTM слоя, които компресират входните данни в тясно четиримерно латентно представяне (bottleneck). След възстановяване на размерността чрез RepeatVector, декодер с два LSTM слоя реконструира първоначалния сигнал, а изходен плътно-свързан слой реализира крайната реконструкция. Моделът поддържа краткосрочна и дългосрочна памет, което позволява ефективно моделиране на времеви зависимости. Допълнително са приложени и алгоритъмът One-Class SVM, който изгражда отделяща хиперплоскост, както и статистическите методи IQR и Z-score.

При тренирането и тестването на моделите е използван телекомуникационният компонент на мащабен набор от данни за град Милано и провинция Трентино. Географската територия на тези две области е разделена на мрежи, съответстващи на квадрати с размери около 235 x 235 метра, като Милано се състои от 1000 квадрата, а Трентино от 6575. “Лабораторията за семантика и иновации в знанието” на Telecom Italia предоставя записите за подробностите на разговорите (Call Detail Records; CDR), чрез които са измерени стойностите на трафика. Всеки път, когато даден потребител взаимодейства с мрежата, се създава нов CDR запис, съдържащ времето на взаимодействието и обслужващата базова станция. Географското местоположение на потребителя се определя чрез картите на покритие, които показват обслужваната територия от всяка базова станция, като взаимодействията се агрегират според квадрата от мрежата, към който принадлежат. Записите са временно

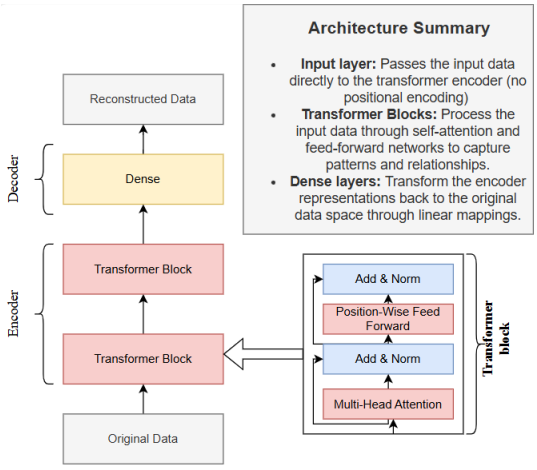
агрегирани във времеви интервали от десет минути и са умножени по константа, определена от оператора, за да се скрие истинският брой на обажданията. Наборът от данни предоставя георефернцирани измервания за двумесечен период от 1 ноември 2013 г. до 1 януари 2014 г. Характеристиките включват идентификатор на квадрата, времеви интервал, както и брой входящи и изходящи SMS съобщения, телефонни обаждания и ниво на интернет трафика. Първоначалните данни се състоят от 62 текстови файла, които са обединени и агрегирани така, че всяка проба да представлява измерване на всеки час. За намаляване на изчислителната сложност в ранните етапи на разработването, наборът от данни е ограничен до 6 от 10-те базови станции с най-високо интернет потребление, като несъществените характеристики са премахнати. Поради несъответствия в измерванията след 22 декември, вероятно породени от празничния сезон, тези анормални данни са изключени. Окончателният набор от данни, състоящ се от 8928 записа, е сортиран хронологично и разделен: наборът за обучение включва данни до 11 декември 2013 г., а тестовият набор обхваща останалите измервания до 22 декември. За включване на времева информация са извлечени характеристики като час, ден от седмицата, ден от месеца и месец, след което данните са стандартизирани и нормализирани.

Anomaly type	GridID	Date	Time	Anomaly label
Internet spike	5059	14.12.2013	10am-8pm	1
SMSIn drop	All grids	18.12.2013	10am-8pm	2
CallOut drop	All grids	16.12.2013	10am-8pm	3

Таблица 5. 1. Описание на изкуствено добавените аномалии

Тъй като оригиналният набор от данни се състои от нормални профили на трафика, първоначално на всички измервания е присвоен етикет за аномалия 0. За да се тестват моделите, в тестовия набор са инжектирани изкуствени аномалии, симулиращи реални проблеми в мрежата. Въведен е скок в интернет потреблението, симулиращ DDoS атака, както и внезапни спадове във входящите SMS съобщения и изходящите гласови обаждания, симулиращи мрежова повреда. Тези 143 аномалии са стратегически въведени по време на пиковите часове между 10:00 и 20:00 ч., когато мрежовият трафик е най-висок.

Архитектурата на предложения Трансформаторен модел използва само блока на енкодера от оригиналната архитектура. Обучавайки се единствено върху нормални данни, моделът научава тяхното разпределение и впоследствие открива аномалии чрез изчисляване на функцията на загуба между реконструираните и оригиналните данни. Архитектурата се състои от енкодер с два слоя, като първият създава скрити представяния, които се подават към втория слой за изграждане на представяния от по-високо ниво. Всеки трансформаторен блок включва механизъм за паралелно самонаблюдение и позиционно-свързана невронна мрежа за предаване напред, оборудвани с остатъчни връзки и нормализация на слоя. Процесът на откриване включва обучение върху чисти времеви редове, определяне на праг за грешката при реконструкция и маркиране на всяко отклонение над този праг като аномалия при тестване с нови данни.



Фигура 5. 5. Архитектура на трансформатора

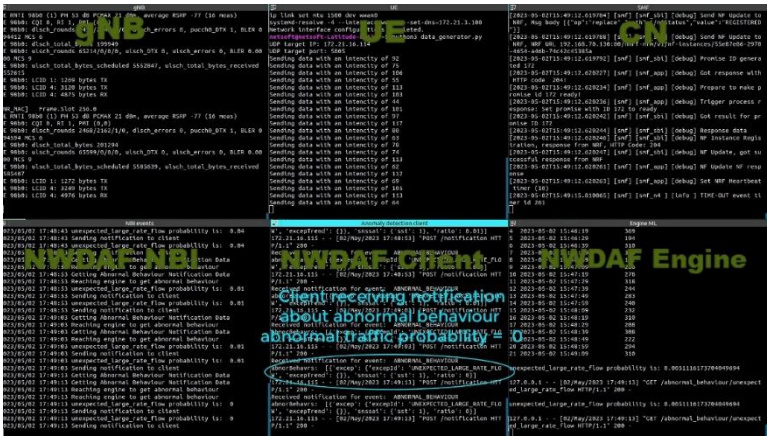
При оценката на резултатите Трансформаторът демонстрира изключителна реконструкционна производителност. След обучение в продължение на оптималните 23 епохи за избягване на пренапасване (overfitting), моделът успешно открива 138 от 143-те инжектирани аномалии, постигайки точност от 96.5% при време за изпълнение от едва 29.45 секунди. За сравнение, LSTM-Autoencoder открива 133 аномалии, постигайки 93.01% точност, но изисква 40 епохи на обучение и значително повече време (77.13 секунди) за преодоляване на прекомерното приспособяване. Алгоритъмът One-Class SVM се представя изненадващо добре, идентифицирайки 123 аномалии (86.01% точност) без необходимост от сложно обучение. От друга страна, статистическите методи IQR и Z-score се оказват напълно неефективни, като откриват едва 10 аномалии (6.99% точност), предимно в интернет трафика поради значителното им отклонение от медианните стойности.

Модел	Брой епохи	Брой открити аномалии	Точност	Време за изпълнение
Трансформатор	23	138	96.50%	29.45s
LSTM-Autoencoder	40	133	93.01%	77.13s
OC-SVM	-	123	86.01%	0.266s
Z-Score	-	10	6.99%	0.215s
IQR	-	10	6.99%	0.111s

Таблица 5. 2. Обобщени резултати

с. Имплементация на ML-базиран модел за откриване на аномалии в тестовата O-RAN мрежа

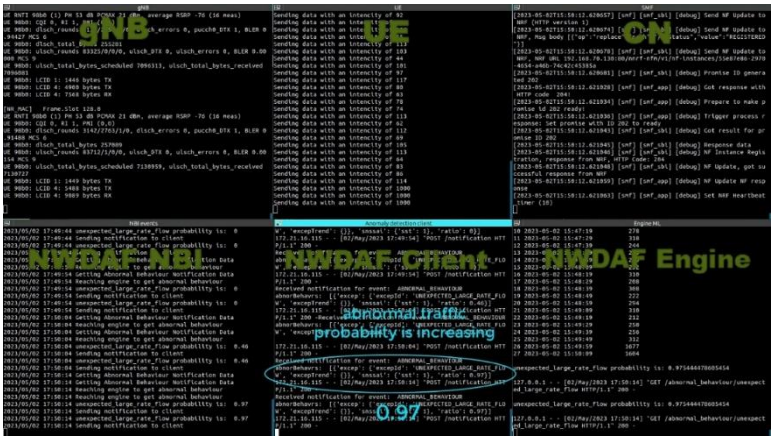
В края на главата се разглежда практическата имплементация на разработения Трансформаторен модел за детекция на аномалии в изградената експериментална O-RAN среда. Алгоритъмът е интегриран като аналитичен компонент в 5G опорната мрежа чрез Функцията за анализ на мрежови данни (NWDAF), която предоставя услуги за оптимизация съгласно спецификациите на 3GPP (Release 17). Използвана е микросървисната имплементация на NWDAF от OpenAirInterface, която е логически разделена на три слоя: слой за излагане (Exposure) за комуникация с външни клиенти и известяване при събития, слой за мониторинг (Monitoring) за събиране на телеметрия от опорната мрежа (Open5GS) и радиомрежата (чрез xApp и near-RT RIC), и Аналитичен слой (Analytics), където като самостоятелна микроуслуга се изпълнява Трансформаторният модел. Работният процес включва непрекъснато извличане на данни за трафика, които се анализират в реално време. Аномалия се регистрира, когато изчислената грешка на реконструкция надхвърли предварително дефиниран праг, след което системата автоматично генерира стандартизирана нотификация за абнормално поведение. Практическата верификация доказва високата надеждност на изградената архитектура.



Фигура 5. 6. Визуализация на нормален интернет трафик. NWDAF отчита липса на аномалии

При сценарий с нормален трафик NWDAF коректно отчита липса на отклонения с минимална грешка на реконструкция. При симулиране на нетипично натоварване чрез изкуствено инжектиран трафик, моделът реагира мигновено, като стойностите на грешката нарастват рязко и само за няколко секунди изчислената вероятност за аномалия достига 97%. Тези резултати категорично потвърждават способността на интегрирания ML модел да

проследява динамиката на мрежата в реална 5G среда и да класифицира ескалацията на аномалиите с изключителна прецизност.

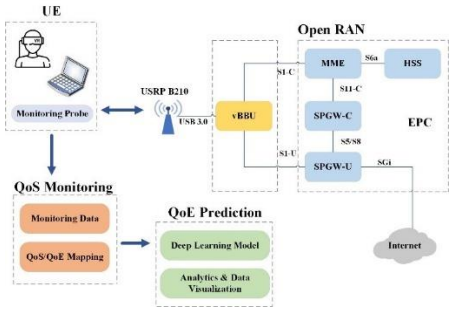


Фигура 5. 7. Класификация на трафика като аномален. NWDAF отчита 97% вероятност за абнормално поведение

6. Модели за оценка и прогнозиране на QoE в интерактивните мултимедийни услуги

Шеста глава от дисертационния труд разглежда фундаменталната необходимост от преход от статична към динамична и прогнозна оценка на качеството на потребителското изживяване (QoE) в съвременните мобилни мрежи. Теоретичното въведение обосновава, че с бързото развитие на интерактивни мултимедийни услуги като облачен гейминг и виртуална реалност (VR), конвенционалните реактивни механизми за мрежова оптимизация вече не са достатъчни. За разлика от традиционното видео, тези услуги са изключително чувствителни към динамиката на радиоканала и изискват комбинация от ултраниска латентност, минимален джитер и липса на загуба на пакети. Дори милисекундна деградация може да доведе до физически дискомфорт при VR потребителите или загуба на управление в гейминга. Поради това управлението на ресурсите в Open RAN архитектурата трябва да бъде проактивно, като мрежата не просто отчита текущото състояние, а предвижда бъдещото поведение на услугата въз основа на исторически данни и текуща телеметрия. В този контекст алгоритмите за машинно обучение се утвърждават като единствения надежден инструмент за моделиране на сложните нелинейни и времеви зависимости между QoS и QoE, преодолявайки ограниченията на класическите аналитични математически модели.

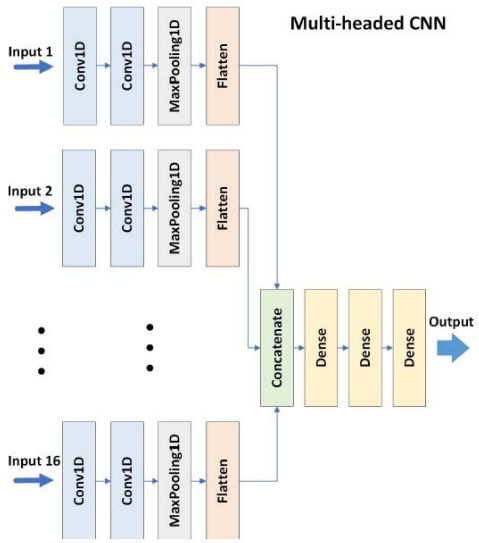
За осигуряване на емпирични данни за обучението и валидацията на прогностичните модели, върху O-RAN тестовата мрежа (описана в Глава 4) е проектирана и интегрирана специализирана система за непрекъснат мониторинг. Системата е реализирана чрез софтуерния стек Prometheus, Telegraf и Grafana и е изцяло ориентирана към потребителя (user-centric), тъй като извлича телеметрия директно от крайните устройства. Събрани са два мащабни набора от данни в реална мобилна среда в продължение на осем седмици, съдържащи по над 80 000 времеви проби. Първият набор обхваща предаване на 4K UHD VR 360-градусови видеа, като регистрира пропускателна способност, закъснение и загуба на пакети. Вторият набор е фокусиран върху стрийминг на 2K гейминг видео с 60 FPS и висока динамика на сцените, като допълнително записва и стойностите на джитера.



Фигура 6. 1. Система за потребителски ориентирано наблюдение на качеството на обслужването (QoS)

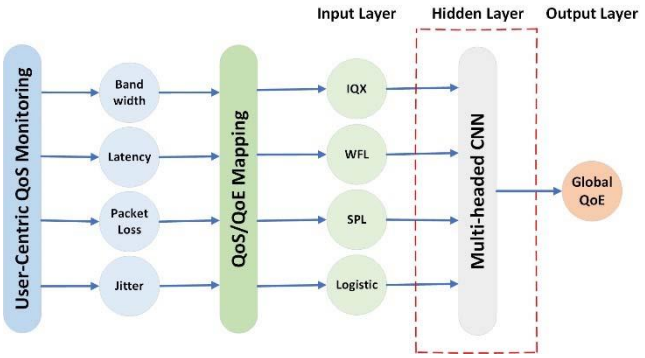
а. Модел за прогнозиране на QoE при стрийминг на гейминг видео

Един от научните приноси в тази глава е разработването на иновативен модел за прогнозиране на QoE при стрийминг на гейминг видео, базиран на архитектура тип Конволюционна невронна мрежа с множество глави (Multi-headed CNN). Моделът е специално проектиран за обработка на многовариантни времеви серии, като разделя входните последователности на паралелни пътища (глави). Всеки път обработва едномерна времева серия на конкретен QoS параметър (закъснение, джитер, загуба на пакети или пропускателна способност). Този подход позволява на модела да улавя специфични локални аномалии, като внезапни пикове в закъснението, които биха се слели и изгубили при използването на стандартна CNN с общ вход. След филтрирането в паралелните пътища, извлечените характеристики се обединяват в общ вектор и се подават към напълно свързани слоеве (dense layers), които моделират глобалните зависимости и генерират крайната прогнозна MOS (Mean Opinion Score) оценка за възприеманото качество.



Фигура 6. 2. Архитектура на Multi-headed CNN

Критичен етап преди подаването на данните към невронната мрежа е предварителното трансформиране на суровите QoS метрики в съпоставими QoE стойности. Тъй като възприеманото качество не се променя линейно спрямо мрежовите параметри, а демонстрира ясно изразени прагови ефекти, в модела са интегрирани нелинейни съпоставящи функции. Използването на логистични криви и експоненциалната хипотеза IQX (в комбинация с психофизичните закони на Фехнер и Стивънс) осигурява прецизно математическо картографиране, което улеснява процеса на обучение на невронната мрежа. Моделът използва исторически прозорец от 48 времеви стъпки (два дни), за да прогнозира средното мнение за качеството през следващите 24 времеви стъпки.



Фигура 6. 3. Преобразуване на QoS към QoE (MOS)

За оценка на ефективността на разработения модел е проведен задълбочен сравнителен анализ срещу други утвърдени архитектури за дълбоко обучение, включително стандартни CNN, LSTM, двупосочни LSTM (Bidirectional-LSTM) и хибридни ResCNN-LSTM мрежи. Емпиричните резултати категорично доказват превъзходството на предложената Multi-headed CNN архитектура. Моделът постига най-ниски нива на грешка по всички ключови статистически метрики, като регистрира Средна абсолютна грешка (MAE) от 0.09786 и Средна абсолютна процентна грешка (MAPE) от едва 2.52%, докато конкурентните LSTM и стандартни CNN модели показват грешки в порядъка на 3.8% - 4.0%. Тези резултати потвърждават, че паралелната конволюционна

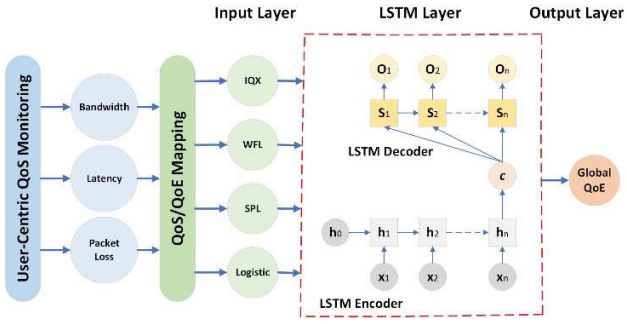
структура е изключително надежден инструмент за прогнозиране на QoE в реални O-RAN мрежи, тъй като успешно съчетава локална чувствителност към кратковременни мрежови флуктуации с глобална обобщаваща способност, правейки модела идеален за интеграция в интелигентните контролери за управление на радиоресурсите в реално време.

	MSE	RMSE	MAE	MAPE (%)	MedAE
CNN	0.03241	0.18005	0.14492	3.92163	0.13820
Multi-channel CNN	0.03432	0.18527	0.14835	4.03715	0.13218
Multi-headed CNN	0.01342	0.11588	0.09786	2.52839	0.09664
TCN	0.03581	0.18924	0.15303	4.14064	0.13052
LSTM	0.03250	0.18029	0.14196	3.85048	0.13774
ResCNN-LSTM	0.03245	0.18016	0.14317	3.87785	0.13492
RNN	0.03450	0.18576	0.14212	3.88562	0.11938
GRU	0.03232	0.17978	0.14060	3.82902	0.12912

Таблица 6. 1. Оценка на модела за прогнозиране на качеството на преживяването (QoE)

b. Модел за прогнозиране на QoE при VR 360-градусово видео

Прогнозирането на качеството на потребителското изживяване (QoE) при VR 360-градусово видео представлява значително по-сложно предизвикателство в сравнение с гейминг видео стрийминга, тъй като виртуалната реалност налага още по-екстремни изисквания към мрежата. За да се гарантира устойчиво имерсивно изживяване, приложенията изискват стабилна и висока пропускателна способност, минимална латентност от край до край, нисък джитер и почти нулева загуба на пакети. Виртуалната реалност е силно чувствителна към моментни аномалии, които могат да предизвикат деградация на визуалното качество, забавяне в реакциите, намаляване на резолюцията или физически дискомфорт (cybersickness), което я позиционира сред най-критичните мултимедийни услуги. Поради тази комплексност, класическите методи за статично QoS/QoE съпоставяне се оказват напълно недостатъчни, тъй като не могат да опишат натрупващите се с времето ефекти върху възприятието. За прецизно прогнозиране в този контекст е разработен иновативен модел, базиран на архитектура тип LSTM encoder–decoder, който е специализиран в обработката на дълги времеви последователности и улавянето на контекстуалните зависимости в структурата на мрежовата деградация.



Фигура 6. 4. Архитектура на модел за прогнозиране на качеството на преживяването чрез дълбоко обучение

Предложеният модел функционира чрез два взаимосвързани компонента – енкодер (encoder) и декодер (decoder). Енкодерът приема като вход последователност от QoS параметри, измерени през предходните 48 времеви стъпки, и ги обработва чрез слоеве от клетки с дълга и краткосрочна памет (LSTM). Основната цел на този компонент е да компресира историческата информация в компактно и богато латентно пространство, което отразява динамичната структура на времевите зависимости, включително предходни тенденции, плавни преходи и признаци за потенциално претоварване на мрежата. Декодерът от своя страна използва това латентно представяне, за да екстраполира зависимостите и да генерира прогноза за бъдещите QoE стойности с хоризонт до 24 стъпки напред (едно денонощие). Това позволява на интелигентните мрежови контролери да реагират проактивно преди реалното влошаване на качеството. Входните данни към модела се генерират чрез предварително нелинейно преобразуване на QoS параметрите в съответни QoE еквиваленти, като латентното представяне служи като своеобразна „памет“ за проследяване на цялостната времева структура на трафика.

За разлика от Multi-headed CNN архитектурата, приложена успешно при гейминг сценариите, LSTM моделът е фокусиран върху дългосрочните зависимости. Докато конволюционните мрежи са изключително ефективни при изолиране на локални вариации, VR съдържанието изисква разпознаване на фини времеви структури, простиращи се през множество проби. Например постепенно увеличаващ се джитер или леко завишена загуба на пакети, които предвещават сериозен срив. Архитектурата постига това чрез механизмите за „запомняне“ и „забравяне“ на информация, вградени във вратите (gates) на LSTM клетките, което позволява на невронната мрежа динамично да определя тежестта на различните исторически моменти.

	MSE	RMSE	MAE	MAPE (%)	MedAE
Naive	0.29636	0.54439	0.49610	15.2277	0.46343
Simple RNN	0.04126	0.20312	0.17529	4.68265	0.17242
LSTM	0.03383	0.18395	0.15186	4.08651	0.13727
Autoencoder LSTM	0.03134	0.17704	0.14987	4.05443	0.14518
Bidirectional LSTM	0.029015	0.17033	0.13958	3.81615	0.12350
Encoder-Decoder LSTM	0.02541	0.15943	0.12491	3.42698	0.09981
GRU	0.02767	0.16361	0.13198	3.63324	0.10320

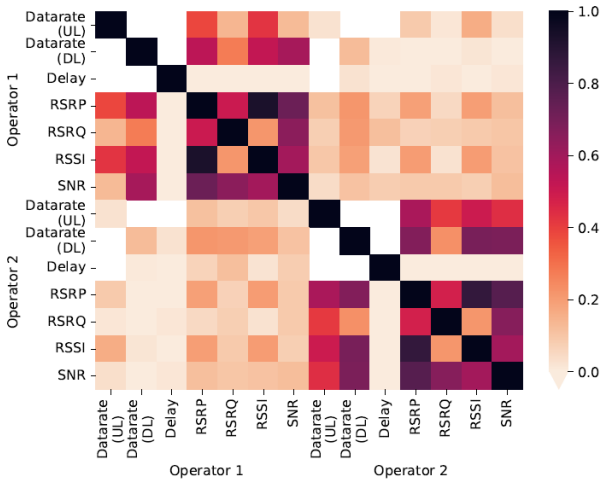
Таблица 6. 2. Оценка на модела за прогнозиране на качеството на преживяването (QoE)

Процесът на обучение е реализиран чрез оптимизиране на средноквадратичната грешка (MSE) върху мащабния набор от реални данни, събрани от VR тестовете в O-RAN средата. Емпиричните експерименти доказват, че оптималният баланс между прецизност и генерализация се постига при използването на точно два LSTM слоя, като за предотвратяване на свръхнапасването (overfitting) са интегрирани dropout слоеве и механизми за ранно прекратяване на обучението. Резултатите от сравнителния анализ категорично подчертават предимствата на разработената архитектура. LSTM encoder-decoder моделът не само успешно следва общата динамика на QoE за 24 стъпки напред, но и демонстрира най-висока точност сред всички тествани съвременни методи. Моделът постига Средноквадратична грешка (MSE) от 0.02541, Средна абсолютна грешка (MAE) от 0.12491 и Средна абсолютна процентна грешка (MAPE) от едва 3.42%. Тези стойности превъзхождат значително конкурентни архитектури като Simple RNN, стандартни еднослойни LSTM мрежи (MAPE 4.08%), Autoencoder LSTM и двупосочни (Bidirectional) LSTM модели. Резултатите доказват, че предложеният LSTM encoder-decoder модел е високоефективен и надежден инструмент за проактивно управление на ресурсите при критични VR услуги в бъдещите интелигентни мобилни мрежи.

7. Прогнозиране на QoS в C-V2X сценарии чрез машинно обучение

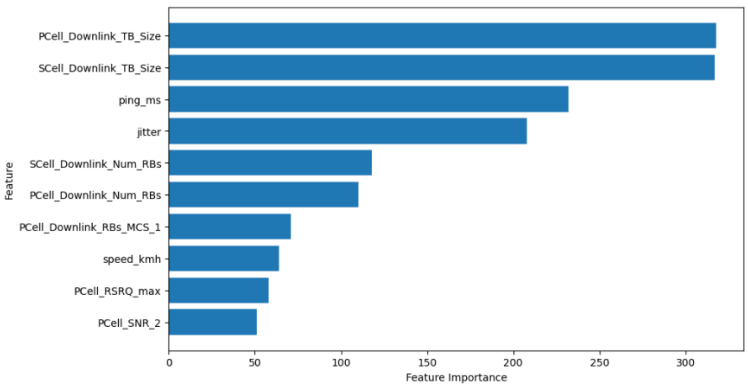
Седмата глава от дисертационния труд разглежда проблематиката на прогнозирането на качеството на услугата (QoS) в сценарии за комуникация от типа „Превозно средство към всичко“ (C-V2X). Теоретичното въведение обосновава критичната роля на C-V2X технологиите за автономното шофиране, където ултранадеждната комуникация с ниско закъснение (URLLC) е въпрос на функционална безопасност, а не просто на потребителски комфорт. В този контекст, способността на мрежата да предоставя прогнозни нива на качеството на връзката (Predictive QoS) е от ключово значение, позволявайки на автономните системи проактивно да адаптират стратегиите си на управление преди реалното влошаване на комуникацията. Въпреки потенциала на машинното обучение, съществуващите изследвания често страдат от липса на реализъм, разчитат прекомерно на GPS координати (което води до запомняне на локалната топология на мрежата) и не демонстрират способност за генерализация между различни мобилни оператори. За да преодолее тези ограничения, главата се фокусира върху разработването и експерименталната валидация на универсални, операторно-независими модели, които прогнозират пропускателната способност в права и обратна посока, разчитайки единствено на параметри от мобилната мрежа, налични на ниво потребителско устройство.

За обучението и валидацията на предложените алгоритми е използван мащабният масив с данни „Berlin V2X Dataset“, събран в реална експлоатационна градска среда. Измерванията обхващат разнообразни сценарии (жилищни зони, паркове, магистрали и тунели) и две отделни мрежи на мобилни оператори, което позволява улавянето на пълната сложност на интерференцията, фадинга и динамичното натоварване на клетките. Корелационният анализ на данните разкрива фундаментални различия в стратегиите на операторите за управление на радиоресурсите. Единият оператор демонстрира агресивно използване на спектъра с висока вариативност, докато вторият прилага по-консервативна политика. Анализът изолира ключовите прогностични признаци, установявайки, че скоростта в права посока (Downlink) корелира силно с физическите параметри на сигнала, най-вече със съотношението сигнал-шум (SNR), докато за връзката в обратна посока (Uplink) определящи са индикаторите за сила на сигнала (RSRP, RSSI). Същевременно се доказва, че закъснението почти не корелира с радиопараметрите, което потвърждава, че латентността зависи предимно от натоварването на използваната инфраструктура.



Фигура 7. 1. Корелационна матрица

Критично архитектурно решение в методологията на предложените модели е пълното изключване на географските координати от входния обучителен вектор. Конструирането на признаците (Feature Engineering) се базира изцяло на комбиниране на информация от физическото ниво (RSRP, RSRQ, SNR) и нивото за управление на достъпа до медията (размер на транспортния блок и схема на модулация и кодиране). Този подход гарантира, че алгоритмите усвояват реалните физически зависимости на радиоразпространението и поведението на мрежовия планировчик, което им позволява да функционират независимо от локацията и да се генерализират успешно към нови мрежови инфраструктури. Анализът на важността на признаците потвърждава, че размерът на транспортния блок, мрежовото закъснение и джитерът играят най-съществена роля за точността на прогнозите, показвайки, че логическото натоварване на клетката е по-критично от чистите физически параметри.



Фигура 7. 2. Топ 10 най-важни параметри

Експерименталната фаза сравнява четири различни класа алгоритми (Линейна регресия, Random Forest, LightGBM и XGBoost), като моделите са обучени върху данните от единия оператор и са валидирани директно върху данните на втория. Резултатите при прогнозирането на пропускателната способност в права посока

(Downlink) категорично доказват превъзходството на дървовидните ансамблови методи над линейните подходи. Алгоритъмът LightGBM демонстрира най-висока обща производителност със стойност на коефициента на детерминация (R^2) от 0.935 и средна абсолютна грешка (MAE) от 3.11 Mbps върху тестовите данни, предлагайки оптимален баланс между точност и изчислителна ефективност. Моделът Random Forest се представя почти идентично (R^2 от 0.934 и MAE 2.98 Mbps), докато линейната регресия претърпява пълен провал с отрицателен R^2 и огромна грешка от 22.4 Mbps, което доказва силно нелинейния характер на процесите в Downlink канала.

Модел	R^2(Train)	R^2(Test)	MAE (Train) [Mbps]	MAE (Test) [Mbps]
LightGBM	0.967	**0.935**	4.29	3.11
Random Forest	**0.994**	0.934	**1.70**	**2.98**
XGBoost	0.980	0.924	3.52	3.31
Linear Regression	0.816	-773.38	12.70	22.40

Таблица 7. 1. Резултати от прогнозиране на пропускателната способност в права посока

При прогнозирането на пропускателната способност в обратна посока (Uplink) динамиката на мрежата се променя, като водещи фактори стават закъснението и джитерът поради ограничения енергиен бюджет на потребителското устройство. В този сценарий Random Forest се очертава като най-прецизния модел, постигайки най-ниска грешка от 1.59 Mbps при R^2 от 0.919, тъй като неговият подход се справя по-успешно със специфичния шум в Uplink данните. LightGBM постига идентичен R^2, но с леко по-висока грешка, а линейната регресия отново се оказва напълно неприложима. Проведените експерименти категорично доказват, че елиминирането на GPS координатите и разчитането единствено на PHY/MAC радиопараметри позволява на ансамбловите нелинейни модели да постигнат над 90% точност дори при тестване в мрежа на напълно различен оператор. Това успешно изпълнява главната цел на изследването – създаването на надежден, операторно-независим прогностичен QoS модел, който е критично необходим за безопасното разгръщане на автономната мобилност в бъдещите комуникационни мрежи.

Модел	R^2(Train)	R^2(Test)	MAE (Train) [Mbps]	MAE (Test) [Mbps]
LightGBM	0.998	0.919	0.878	1.68
Random Forest	**0.998**	**0.919**	**0.33**	**1.59**
XGBoost	0.99	0.88	0.77	2.49
Linear Regression	0.38	-1.73	9.64	14.53

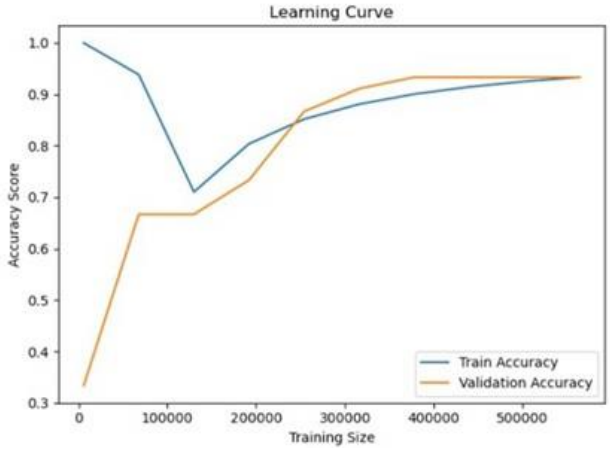
Таблица 7. 2. Резултати от прогнозиране на пропускателната способност в обратна посока

8. Оптимизация на мрежовите ресурси и управление на качеството чрез мрежово нарязване

Осма глава от дисертационния труд разглежда практическото затваряне на цикъла за интелигентно управление в мобилните мрежи, преминавайки от фазите на мониторинг (Глава 5) и прогнозиране (Глави 6 и 7) към етапа на реална оптимизация и изпълнение. Експоненциалното увеличение на мрежовия трафик налага разработването на гъвкави архитектури, способни да управляват динамично споделяне на ресурсите при строг контрол на достъпа. В този контекст концепцията за мрежово нарязване (Network Slicing), въведена от NGMN алианса, предоставя фундаментален механизъм за създаване на множество независими логически мрежи (срезове) върху обща физическа инфраструктура. Всеки срез е изолиран и оптимизиран за специфични изисквания за качество на услугата, което позволява съвместното съществуване на хетерогенни услуги. Управлението на тази архитектура обаче изисква намирането на сложен баланс между персонализирането на услугите, ефективността на ресурсния мениджмънт и общата системна сложност. За преодоляване на тези предизвикателства изследването се фокусира

върху две основни направления: автоматизиран избор на мрежов срез чрез машинно обучение и динамична оптимизация на радиоресурсите в реална O-RAN среда.

За реализирането на автоматизиран избор на мрежови срезове е разработена методология за интелигентна класификация на трафика, която проактивно насочва потребителските заявки към подходящото мрежово парче (eMBB, URLLC или mMTC). Използван е специализираният набор от данни „Deep Slice“, съдържащ ключови показатели, извличани от контролните съобщения между устройството и мрежата, като допустимо забавяне, максимална загуба на пакети и поддържана технология. След предварителна обработка за премахване на излишните характеристики, е проведен анализ чрез Дърво на решенията, който изолира най-важните класификатори. Установено е, че процентът на загуба на пакети под 0.001 ефективно разграничава критичния URLLC трафик, поддържаната технология разделя eMBB от mMTC, а бюджетът за закъснение прецизира границата между mMTC и URLLC. Първоначалното обучение демонстрира 100% точност, което обаче е индикатор за силно пренапасване (overfitting) поради ограничения брой уникални записи в суровия набор, което би компрометирало генерализацията в реална среда.



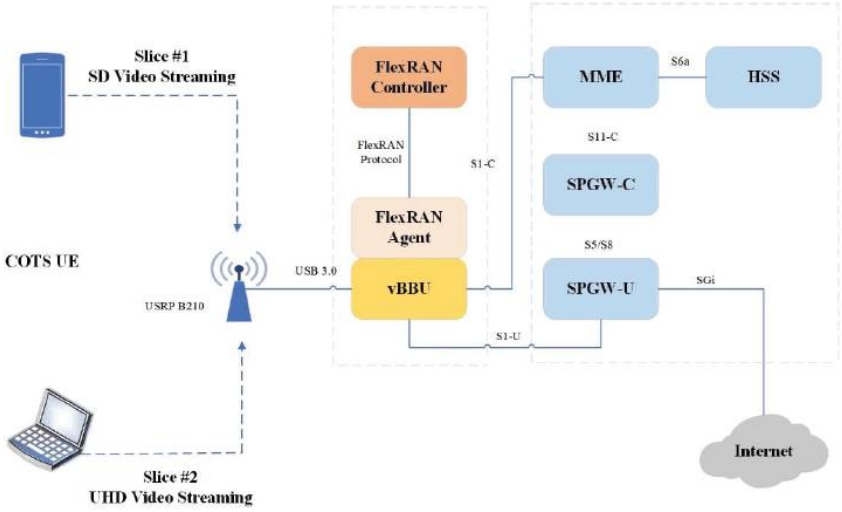
Фигура 8. 1. Крива на обучението след свръхдискретизация (oversampling) и дълбоко конструиране на признаците (feature engineering)

За да се преодолее проблемът с генерализацията и да се създаде устойчив модел, е приложен процес на дълбоко конструиране на признаците (feature engineering). Всички входни характеристики са преобразувани в бинарен формат, което първоначално понижава точността до 91%, но драстично подобрява паралелизма между кривите на обучение и валидиране, гарантирайки адекватна работа с непознати данни. Допълнително предизвикателство представлява сериозният класов дисбаланс, при който URLLC трафикът доминира в почти половината от записите. Този проблем е решен чрез прилагане на техника за свръхдискретизация (oversampling), което балансира класовете и повишава точността до 92%. За финален класификатор е избран ансамбловият алгоритъм AdaBoostClassifier, който използва последователно свързани дървета на решенията за минимизиране на грешките. Моделът е изключително бърз и изчислително лек, което го прави идеален за имплементация като xApp приложение в периферията на мрежата. След обучение върху 80% от данните, финалният модел постига точност от 93.4%, като матрицата на объркванията доказва минимален и статистически несъществен брой фалшиво-положителни грешки, потвърждавайки успешната класификация на трафика.

	URRLC	eMBB	mMTC
Точност (Accuracy)	0,934		
Прецизност (Precission)	1	1	0,834
Пълнота (Recall)	0.801	1	1
F-Оценка (F-Score)	0.89	1	0,91

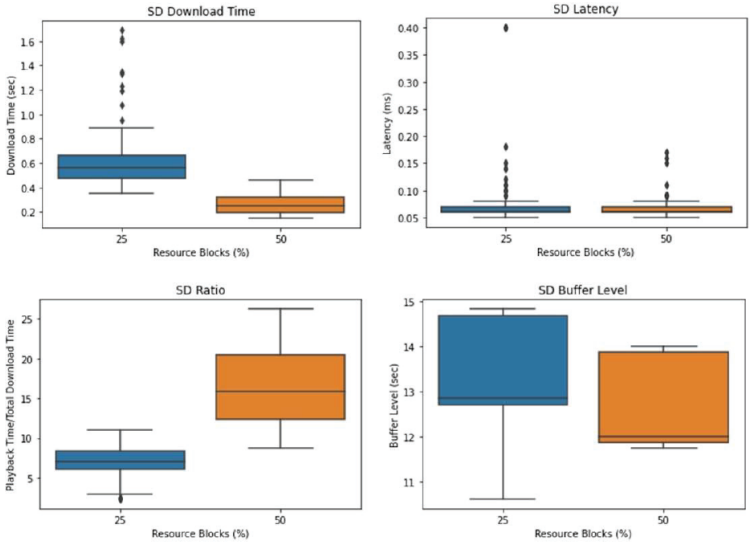
Таблица 8. 1. Ключови показатели за ефективност на модела

Второто направление от изследването надгражда успешната класификация чрез практическа демонстрация на динамично разпределение на ресурсите (PRBs) в реална O-RAN тестова мрежа с цел гарантиране на QoS при мултимедийни услуги. Експерименталната постановка използва платформата с отворен код FlexRAN в ролята на Near-RT RIC и интегрира специализирана система за мониторинг на DASH видео стрийминг.



Фигура 8. 2. Експериментална среда за Network Slicing

Създадени са два отделни среза - един за видео със стандартна дефиниция (SD) и един за ултрависока дефиниция (UHD). При статично равно разпределение на ресурсните блокове (50% за SD и 50% за UHD), резултатите показват, че ресурсите са напълно достатъчни за SD потока (нула изпуснати кадри и ниска латентност), но са абсолютно недостатъчни за UHD видеото, което търпи критична деградация с недопустими закъснения и 6897 изпуснати кадри.



Фигура 8. 3. SD видео KPIs при различно разпределение на радиоресурсите

За разрешаване на този проблем е разработен автономен скрипт, който осъществява непрекъснат мониторинг на ключовите показатели и при регистриране на деградация в UHD потока генерира тригер към API-то на FlexRAN. Контролерът реагира в реално време, като динамично преконфигурира мрежата – ограничава ресурсите за SD среза до 25% и приоритетно разширява капацитета на UHD среза до 75%. Емпиричните резултати от тази динамична оптимизация са категорични: намаляването на ресурсите за SD видеото оказва минимално, почти незабележимо въздействие (едва 7 изпуснати кадри), докато качеството на UHD стрийминга се подобрява драстично. Времето за изтегляне и латентността спадат значително, нивото на буфера се стабилизира, а броят на изпуснатите кадри намалява над 43 пъти (до 158). Този експеримент категорично доказва, че интеграцията на интелигентен мониторинг с динамично мрежово нарязване в RAN периферията е ключов механизъм за ефективно управление на радиоресурсите и гарантиране на високо качество при критични мултимедийни услуги.

SD (25% PRBs)	SD (50% PRBs)	UHD (50% PRBs)	UHD (75 % PRBs)
7	0	6897	158

Таблица 8. 2. Изгубени кадри при SD и UHD видео стрийминг

Изводи и заключения:

Дисеретационният труд изследва и валидира хипотезата, че интегрирането на изкуствен интелект (AI) и машинно обучение (ML) в архитектурата на Open RAN е не само възможно, но и необходимо за ефективното управление на съвременните и бъдещите мобилни мрежи. Чрез теоретичен анализ, проектиране на комплексна експериментална среда и провеждане на серия от практически експерименти, дисертацията демонстрира, че интелигентните алгоритми могат успешно да решат проблемите със сложността, динамиката и разнородните изисквания на услугите, които традиционните методи за управление не успяват да адресират. В рамките на изследването са постигнати следните ключови научни приноси:

1. **Създаване на реалистична експериментална платформа:** Успешно е проектирана и внедрена пълнофункционална O-RAN базирана тестова мрежа, интегрираща компоненти с отворен код и хардуер за широка употреба. Тази платформа служи като основа за генериране на уникални масиви от реални данни и за валидиране на разработените модели в среда, която отразява реалните хардуерни и софтуерни ограничения на мрежата. Сравнителният анализ на производителността в LTE и 5G режими потвърждава, че архитектурите с функционално разделяне (Functional Splits) предлагат сравнима производителност с традиционните монолитни решения, но с добавена гъвкавост.
2. **Повишаване на сигурността и надеждността чрез AI/ML:** Дисертацията доказа превъзходството на моделите за дълбоко обучение пред класическите статистически методи при откриването на мрежови аномалии. Разработеният и интегриран в OAI NWDAF трансформаторен модел (Transformer) демонстрира способност да открива сложни аномалии в трафика с точност от 96.5% , значително надвишавайки резултатите на методи като Z-score и IQR. Това позволява проактивна реакция в близко до реалното време.
3. **Прогностично управление на QoE за интерактивни услуги:** Разработени са специализирани ML модели за прогнозиране на качеството на потребителското изживяване (QoE) за услуги, силно чувствителни към мрежовите параметри:
    - За **гейминг видео стрийминг** е валидиран Multi-headed CNN модел, който ефективно улавя краткосрочните флуктуации в джитера и латентността, постигайки най-ниска грешка при прогнозиране.
    - За **VR 360-градусово видео** е създаден LSTM Encoder-Decoder модел, способен да моделира дългосрочните времеви зависимости и да предсказва качеството за 24 стъпки напред.
    - За **C-V2X сценарии** е доказано, че локационно-независими модели (като LightGBM и Random Forest), обучавани само върху радио параметри без GPS координати, могат да прогнозират пропускателната способност с висока точност ($R^2 > 0.9$) дори при пренос между мрежите на различни оператори.
4. **Затваряне на цикъла за управление чрез автоматизация:** Дисертацията демонстрира практическата реализация на затворен цикъл на управление (closed-loop control) в O-RAN среда. Разработеният механизъм за динамично мрежово нарязване и преразпределение на ресурсите, базиран на мониторинг в реално време, показва способност да възстанови качеството на UHD видео поток чрез автоматизирано приоритизиране на трафика. Допълнително е валидиран ML модел за автоматична класификация и асоцииране на потребители към правилното мрежово парче с точност над 93%.

Резултатите от дисертационния труд потвърждават, че архитектурата Open RAN предоставя необходимите инструменти и интерфейси за превръщането на мобилната мрежа от статична инфраструктура в интелигентна, адаптивна платформа. Чрез използването на модели за дълбоко обучение за прогнозиране и детекция, комбинирани с програмируемостта на RIC контролерите, е възможно да се постигне ниво на оптимизация и гаранция за качеството на услугите (QoS/QoE), което е непостижимо с традиционните методи. Разработените модели и алгоритми са операторно-независими и приложими в реални сценарии, като предлагат решения на критични проблеми като управлението на трафика за VR и автономни автомобили. Този труд допринася както за теоретичното разбиране на AI-native мрежите, така и за практическото им реализиране чрез предоставяне на валидирани архитектурни решения и софтуерни имплементации.

Литература

[1] G. Heine and H. Sagkob, GPRS: Gateway to Third-Generation Mobile Networks. London, U.K.: Artech House, 2003.

[2] Colonna, Massimo & Barbaresi, Andrea & Zarba, Giovanna & Mantovani, Andrea. (2008). A Brief Survey of VoIP QoS over a multi-RAT Heterogeneous Wireless Network.

[4] Technical Specifications and Technical Reports for UTRAN-Based 3GPP System, 3GPP TR 21.10. 2003

[5] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Release 8 Overview," Release 8, Dec. 2008. [Online]. Available: https://www.3gpp.org

[8] 5G NR Network Interfaces: Xn, NG, E1, F1, F2 Explained [Available at: https://www.rfwireless-world.com/tutorials/5g/5g-nr-network-interfaces] [Last Accessed: 22.05.2025]

[17] Study on new radio access technology Radio access architecture and interfaces, 3GPP, 3rd Generation Partnership Project, 3 2017, v14.0.

[21] Xu, F., Yao, H., Zhao, C. et al. Towards next generation software-defined radio access network–architecture, deployment, and use case. J Wireless Com Network 2016, 264 (2016). https://doi.org/10.1186/s13638-016-0762-6

[30] M. Polese, L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," in IEEE Communications Surveys & Tutorials, vol. 25, no. 2, pp. 1376-1411, Secondquarter 2023, doi: 10.1109/COMST.2023.3239220.

[35] S. Suthaharan, "Supervised learning algorithms," in Machine Learning Models and Algorithms for Big Data Classification, New York, NY, USA: Springer, 2016, p. 183–206.

[36] Tyagi, Kanishka & Rane, Chinmay & Sriram, Raghavendra & Manry, Michael. (2022). Unsupervised learning. 10.1016/B978-0-12-824054-0.00012-5.

[37] A. Paz and S. Moran, "Non deterministic polynomial optimization problems and their approximations," Theoretical Computer, vol. 15, no. 3, p. 251–277, 1981.

[42] O-RAN AI/ML Workflow Architecture and Framework - ERK, https://erk.fe.uni-lj.si/2024/papers/cop(o_ran_ai_ml).pdf, [Last Accessed: 05.07.2025]

[52] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis and P. I. Lazaridis, "A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction," in IEEE Access, vol. 10, pp. 19507-19538, 2022, doi: 10.1109/ACCESS.2022.3149592.

[57] Canadian Radio-television and Telecommunications Commission (CRTC), "Telecom Decision CRTC 2018-241: CISC Network Working Group – Non-consensus report on quality of service metrics to define high-quality fixed broadband Internet access service," Ottawa, Canada, 13 July 2018. [Online]. Последно достъпено на 13.07.2025 Линк: https://crtc.gc.ca/eng/archive/2018/2018-241.htm

[61]https://gamerhub.co.uk/gaming-industry-dominates-as-the-highest-grossing-entertainment-industry/ Достъпен на: 17.07.2026

[71] Cross-layer latency analysis for 5G NR in V2X communicationsHorta J, Siller M, Villarreal-Reyes S (2025) Cross-layer latency analysis for 5G NR in V2X communications. PLOS ONE 20(1): e0313772. https://doi.org/10.1371/journal.pone.0313772

[56] Alreshoodi, Mohammed & Woods, John. (2013). Survey on QoE\QoS Correlation Models For Multimedia Services. International Journal of Distributed and Parallel systems. 4. 10.5121/ijdps.2013.4305.

[91] Joe Breen, Andrew Buffmire, Jonathon Duerig, Kevin Dutt, Eric Eide, Mike Hibler, David Johnson, Sneha Kumar Kasera, Earl Lewis, Dustin Maas, Alex Orange, Neal Patwari, Daniel Reading, Robert Ricci, David Schurig, Leigh B. Stoller, Jacobus Van der Merwe, Kirk Webb, and Gary Wong. 2020. POWDER: Platform for Open Wireless Data-driven Experimental Research. In Proceedings of the 14th International Workshop on Wireless Network Testbeds, Experimental

evaluation & Characterization (WiNTECH '20). Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/3411276.3412204

[92] Platforms for Advanced Wireless Research Link: https://advancedwireless.org/, Достъпено на 27.07.2025г

[124] Alamr, Abrar, and Abdelmonim Artoli. 2023. "Unsupervised Transformer-Based Anomaly Detection in ECG Signals" Algorithms 16, no. 3: 152. https://doi.org/10.3390/a16030152

Списък на публикациите по дисертационния труд

[A1] Vlahov, Atanas & Ekova, Dessislava & Poulkov, Vladimir & Cooklev, Todor. (2022). Virtualized, Open and Intelligent: The Evolution of the Radio Access Network. 10.1201/9781003360889-9.

[A2] Velyova, Vesela & Vlahov, Atanas & Poulkov, Vladimir & Ivanov, Antoni. (2024). O-RAN Based User Tracking for Emergency Scenarios. 1-5. 10.1109/WPMC63271.2024.10863025.

[A3] Kougioumtzidis, Georgios & Vlahov, Atanas & Poulkov, Vladimir & Zaharis, Zaharias & Lazaridis, Pavlos. (2022). QoE-Oriented Open Radio Access Networks for Virtual Reality Applications. 491-496. 10.1109/WPMC55625.2022.10014946.

[A4] Vlahov, Atanas & Poulkov, Vladimir & Mihovska, Albena. (2021). Analysis of Open RAN Performance Indicators Related to Holographic Telepresence Communications. 1-5. 10.1109/WPMC52694.2021.9700477.

[A5] Mihovska, Albena & Vlahov, Atanas & Poulkov, Vladimir. (2024). 6G-based Intelligent, Context-Aware, and Trustworthy User-Centric Healthcare Applications. 1-6. 10.1109/WTS60164.2024.10536689.

[A6] Evgenieva, Evgeniya & Vlahov, Atanas & Ivanov, Antoni & Poulkov, Vladimir & Manolova, Agata. (2025). A Comprehensive Survey of 6G Simulators: Comparison, Integration, and Future Directions. Electronics. 14. 3313. 10.3390/electronics14163313.

[A7] A. Vlahov, V. Poulkov, P. Lazaridis and Z. Zaharis, "A Machine Learning Methodology for Network Anomalies Detection in O-RAN Networks," European Wireless 2023; 28th European Wireless Conference, Rome, Italy, 2023, pp. 174-178.

[A8] Georgieva, Polya & Vlahov, Atanas & Mfondoum, Roland & Poulkov, Vladimir & Zaharis, Zaharias. (2025). Informer-Based Anomaly Detection in Mobile Networks Using CDR Time-Series Analysis. 1-4. 10.1109/ICEST66328.2025.11098317.

[A9] Gotseva, Nikol & Vlahov, Atanas & Mfondoum, Roland & Ivanov, Antoni & Poulkov, Vladimir. (2025). A Comparative Analysis of Anomaly Detection Techniques in Cellular Data. 1-5. 10.1109/ICEST66328.2025.11098230.

[A10] Kougioumtzidis, Georgios & Vlahov, Atanas & Poulkov, Vladimir & Lazaridis, Pavlos & Zaharis, Zaharias. (2024). QoE Prediction for Gaming Video Streaming in O-RAN Using Convolutional Neural Networks. IEEE Open Journal of the Communications Society. PP. 1-1. 10.1109/OJCOMS.2024.3362275.

[A11] Kougioumtzidis, Georgios & Vlahov, Atanas & Poulkov, Vladimir & Lazaridis, Pavlos & Zaharis, Zaharias. (2023). Deep Learning-Aided QoE Prediction for Virtual Reality Applications Over Open Radio Access Networks. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3343846.

[A12] Gotseva, Nikol & Vlahov, Atanas & Poulkov, Vladimir & Manolova, Agata. (2024). ML-Driven Prediction of QoS in C-V2X Scenarios. 1-4. 10.1109/ICEST62335.2024.10639683.

[A13] Georgieva, Polya & Vlahov, Atanas & Poulkov, Vladimir & Manolova, Agata. (2024). A Machine Learning Approach for Network Slice Selection. 1-5. 10.1109/ICEST62335.2024.10639750.

[A14] Vlahov, Atanas & Kougioumtzidis, Georgios & Mihovska, Albena & Poulkov, Vladimir. (2022). Performance Analysis of Evolved RAN Architectures with Open Interfaces. Journal of Mobile Multimedia. 10.13052/jmm1550-4646.19112.

TECHNICAL UNIVERSITY OF SOFIA
FACULTY OF TELECOMMUNICATIONS
DEPARTMENT "COMMUNICATION NETWORKS"

Atanas Vlahov, M.Sc.

Intelligent management of access networks with open interfaces for the implementation of QoS-critical services

ABSTRACT of Ph.D. THESIS

Future cellular networks are expected to support the diversified requirements of emerging services, with a strong focus on applications that are highly critical to Quality of Service (QoS) and Quality of Experience (QoE), such as autonomous mobility, industrial automation, and highly immersive multimedia. This broad and heterogeneous spectrum of scenarios necessitates the development of flexible, scalable, and programmable networks capable of guaranteeing ultra-high reliability and low latency, ensuring uncompromising quality of service and quality of experience levels for end-users in a dynamic environment.

The main objective of this dissertation is to propose a network management methodology specifically targeted at QoS-critical applications. The implementation is based on leveraging the advantages of machine learning and the Open Radio Access Network (Open RAN) architecture to improve network efficiency, flexibility, and automation. More specifically, the proposed framework focuses on the development of artificial intelligence algorithms integrated into the RAN Intelligent Controller (RIC) to provide autonomous radio resource allocation, precise Network Slicing, as well as continuous QoS monitoring and forecasting. This ensures strict adherence to the network requirements of critical applications and prevents service degradation in real time.

TECHNICAL UNIVERSITY OF SOFIA
FACULTY OF TELECOMMUNICATIONS
DEPARTMENT “COMMUNICATION NETWORKS”

Atanas Vlahov, M.Sc.

INTELLIGENT MANAGEMENT OF ACCESS NETWORKS WITH OPEN INTERFACES FOR THE IMPLEMENTATION OF QOS-CRITICAL SERVICES

EXTENDED ABSTRACT OF A DISSERTATION FOR THE AWARD OF THE EDUCATIONAL AND SCIENTIFIC DEGREE "DOCTOR" (PHD)

Field of higher education: 5. Technical Sciences

Professional field: 5.3 Communication and Computer Engineering

Scientific specialty: Communication Networks and Systems

Scientific supervisor: **Prof. Vladimir Kostadinov Pulkov, DSc, Eng.**

София, 2026

The dissertation was discussed and recommended for defense by the Department Council of the "Communication Networks" Department at the Faculty of Telecommunications, Technical University of Sofia, at a regular meeting held on April 14, 2026 (Protocol No. 9).

The public defense of the dissertation will take place on July 7, 2026, at 3:00 PM in the Conference Hall of the Library and Information Center (LIC) of the Technical University of Sofia at an open meeting of the Scientific Jury, appointed by Order No. ОЖ-5.3-36 / 04.05.2026 of the Rector of TU-Sofia, consisting of:

6. Prof. Georgi Iliev, PhD – Chairman
7. Assoc. Prof. Georgi Balabanov, PhD - Scientific Secretary
8. Prof. Stanimir Sadinov, PhD
9. Prof. Rozalina Dimova, PhD
10. Prof. Gabriela Atanasova, PhD

Reviewers:

1. Prof. Georgi Iliev, PhD
2. Prof. Gabriela Atanasova, PhD

The related materials are available for the interested parties at the office of the Faculty of Telecommunications, Building 1, Room 1439-B, and on the website of the Technical University of Sofia.

The PhD candidate is a part-time doctoral student at the "Communication Networks" Department, Faculty of Telecommunications. The research for the dissertation was conducted by the author, and its results have been published.

Author: Atanas Vlahov, M.Sc.

Title: Intelligent management of access networks with open interfaces for the implementation of QoS-critical services

Тираж: 30 copies

Printed at the Publishing and Printing Complex of the Technical University of Sofia

III. GENERAL CHARACTERISTICS OF THE DISSERTATION

Relevance of the problem

Over the last decade, the telecommunications industry has been undergoing a fundamental transformation, driven by the exponential growth of mobile traffic and the emergence of new services. The deployment of fifth-generation (5G) and the preparation for sixth-generation (6G) mobile networks require the support of diverse use cases: from massive Machine-Type Communications (mMTC) to Ultra-Reliable Low-Latency Communications (URLLC).

Applications such as autonomous mobility, industrial automation, virtual and augmented reality (VR/AR), and holographic telepresence impose extremely strict requirements on Quality of Service (QoS) and Quality of Experience (QoE). Traditional Radio Access Network (RAN) architectures are monolithic, dependent on specialized hardware, and feature closed interfaces, which limits their flexibility and slows down innovation.

In response to these limitations, the concept of an access network with open interfaces (Open RAN) has emerged, introducing the principles of disaggregation and virtualization. However, the dynamic nature of the radio environment and the heterogeneity of services render traditional management methods ineffective, necessitating the integration of Artificial Intelligence (AI) and Machine Learning (ML) to achieve real-time automation and dynamic allocation of network resources.

Objective of the dissertation, main tasks, and research methods

The objective of the dissertation is to propose a comprehensive methodology for improving QoS and QoE by integrating Artificial Intelligence (AI) into network operations to optimize their performance. AI-driven network operations are integrated into the Open RAN architecture to support various use cases with heterogeneous QoS/QoE requirements in terms of bandwidth, latency, packet loss, and jitter.

Specifically, Machine Learning (ML) algorithms are developed to provide automated network traffic prediction and optimal resource allocation utilizing Network Slicing techniques. The implementation of ML-assisted resource orchestration automation leads to optimal network management and provide a new and effective methodology for reducing the Capital Expenditures (CAPEX) and Operational Expenditures (OPEX) of Communication Service Providers (CSPs). To achieve the stated objective, the following tasks have been defined:

6. Analysis and systematization of the RAN architectural evolution and the transition toward open and virtualized architectures (C-RAN, vRAN, O-RAN), as well as the role of RAN Intelligent Controllers (RIC).
7. Design and implementation of a fully functional experimental O-RAN testbed, based on open-source software and COTS (Commercial Off-The-Shelf) hardware.
8. Development of anomaly detection algorithms using deep learning models for the proactive identification of atypical behavior in network traffic.
9. Modeling and forecasting of QoS and QoE in interactive multimedia services (gaming, VR) and C-V2X (Cellular Vehicle-to-Everything) communications through machine learning..
10. Optimization of network resources through mechanisms for intelligent traffic classification and dynamic radio resource allocation.

Original Scientific Contributions (Scientific Novelty)

The scientific novelty of the dissertation consists of:

- Proposing a network management methodology specifically targeted at QoS-critical applications through the integration of AI algorithms into the RAN Intelligent Controller (RIC).
- Developing and implementing an innovative Transformer model for detecting complex network anomalies, which outperforms traditional methods and LSTM architectures in terms of accuracy and training time.
- Creating specialized predictive models for QoE and QoS:
    - A multi-headed CNN architecture for predicting quality in gaming video streaming
    - An LSTM encoder-decoder model for VR 360-degree video, capturing long-term temporal dependencies.
    - A location-agnostic approach for QoS prediction in C-V2X scenarios, allowing a high degree of generalization across different mobile network operators
- **Successful implementation of an O-RAN-based methodology** for the adaptive and dynamic allocation of Physical Resource Blocks (PRBs) among network slices, ensuring guaranteed quality for UHD video streams.

Practical applicability

All developed methods and algorithms, as well as the proposed improvements to already implemented ones, have been investigated and analyzed through simulation experiments. A comparison has also been made with other existing models that are based on similar functionalities and characteristics, or that share comparable objectives regarding the improvement of their operational parameters. All of this makes the potential implementation of the results from this dissertation easily realizable in modern telecommunication networks.

Additional evidence of this high applicability is the successful integration of the proposed machine learning algorithms as microservices into real network components, such as the Network Data Analytics Function (NWDAF) and the RAN Intelligent Controllers, which categorically confirms their interoperability and readiness for practical application.

Publication of the dissertation results

The conducted analyses, proposed approaches, and obtained results for the period 2021–2025 are presented in a total of 14 authored publications indexed in Scopus and Web of Science: 1 book chapter; 9 publications in international conferences; 4 publications in international scientific journals ranked Q1 and Q2. The articles have a total of 59 citations in Scopus and 64 citations in Google Scholar.

Structure and volume of the dissertation

The dissertation is written in Bulgarian and consists of **193** A4 pages. It contains an introduction, eight chapters, a conclusion outlining the main contributions, a list of figures, a list of tables, a list of abbreviations, a list of publications related to the dissertation, a list of references, and two appendices. The main body of the dissertation contains **70 figures** and **13 tables**. A total of **138 references** are

used, all of which are in the Latin script, and over **80%** are from the last ten years. The numbering of the figures and tables in the extended abstract corresponds to that in the dissertation.

.

IV. CONTENT OF THE DISSERTATION

9. Evolution of radio access networks and their architectures

The first chapter of the dissertation presents a comprehensive analysis of the historical development and architectural evolution of radio access networks, tracing the path from early digital systems to modern open and disaggregated environments. The exposition begins with an examination of second-generation architectures, explaining the structure of the base station subsystem in GSM and GPRS [1]. It is pointed out that the introduction of packet switching through GERAN [2] marks an important milestone, but the architecture remains limited due to its monolithic nature and dependence on specific hardware. In the analysis of the third generation, the role of UTRAN [4] is examined, where radio resource management becomes more complex due to the use of WCDMA, requiring new mechanisms for power and mobility control.

Special attention in the chapter is given to the transition to fourth-generation LTE and its access network E-UTRAN [5]. Here, the concept of a flat IP architecture is explored in detail, where the functions of the control node are integrated directly into the eNodeB base station. This approach significantly reduces latency and simplifies the network topology, which is a critical step toward supporting modern mobile services. The dissertation traces how this decentralization paves the way for fifth-generation mobile networks, where the gNodeB base station [8] is now viewed as a set of logically disaggregated units. The functional split between centralized and distributed units is discussed, allowing exceptional flexibility in the deployment of network functions depending on the specific application requirements for capacity and latency.

An important part of the exposition is also the analysis of the Cloud RAN [17] and Software-Defined RAN [21] paradigms. The advantages of centralized signal processing in BBU clusters (BBU pools) are investigated, which enables better resource coordination and cost reduction. At the same time, the limitations of these systems are noted, related to the high capacity requirements of the transport network and the use of closed interfaces. This logically justifies the need for an Open RAN architecture, which is defined as an evolutionary step toward full interoperability among different vendors.

The chapter concludes with a detailed technical overview of the O-RAN architecture [30]. The main components are defined: Near-Real-Time RIC, Non-Real-Time RIC, O-CU, and O-DU, explaining their interaction through the open A1, E2, and O1 interfaces. It is emphasized that intelligence is embedded within the network structure itself through these controllers, enabling the deployment of xApps and rApps for dynamic optimization. This detailed architectural breakdown serves as the foundation for the subsequent chapters, which propose specific machine learning algorithms integrated into the described framework. In the conclusion of the first chapter, it is deduced that an open and disaggregated architecture is a mandatory prerequisite for the implementation of services with critical Quality of Service (QoS) requirements in the heterogeneous environment of future networks.

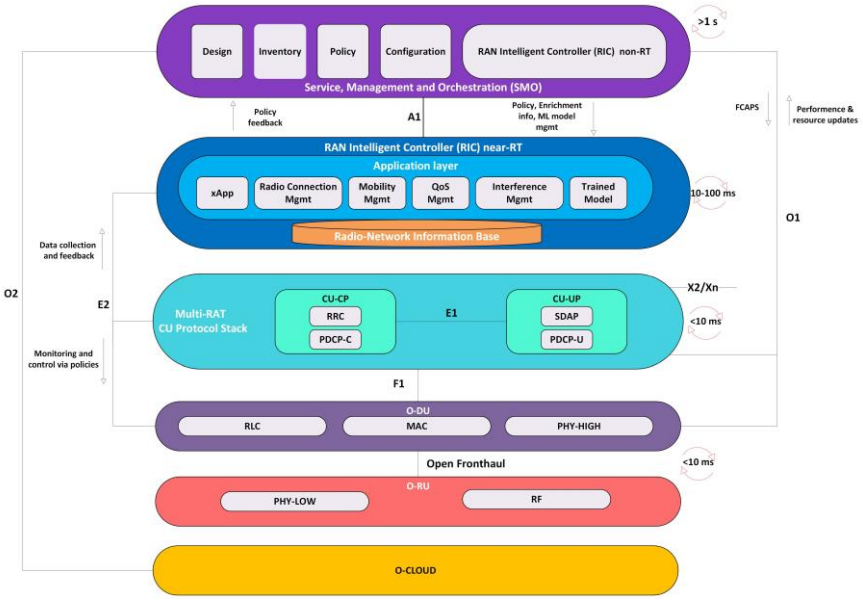


Figure 1.4. O-RAN Architecture

10. Foundations of AI/ML in cellular networks

The second chapter of the dissertation is dedicated to an in-depth investigation of the theoretical foundations and practical aspects of integrating artificial intelligence and machine learning into modern radio access networks. The exposition begins with a detailed taxonomy of machine learning algorithms, providing a critical review of their applicability for the optimization of network operations. The main paradigms of supervised learning [35], unsupervised learning [36], and reinforcement learning [37] are examined, with specific use cases in cellular systems outlined for each. It is emphasized that while supervised learning is indispensable for classification and regression tasks in the presence of labeled historical data, unsupervised and reinforcement learning provide unique capabilities for real-time decision-making under dynamically changing radio conditions..

A major emphasis in the chapter is placed on the machine learning model lifecycle within the O-RAN architecture, in accordance with the specifications of the O-RAN Alliance [43]. The processes of data collection, preprocessing, training, deployment, and monitoring of the models are described in detail. The role of the Non-Real-Time RIC for offline training and policy management is examined, as well as that of the Near-Real-Time RIC for executing trained models in the form of xApps for control within milliseconds. The challenges related to the interoperability of models from different vendors and the need for standardized data exchange interfaces (such as E2 and A1) to ensure closed-loop control and network self-optimization are also discussed.

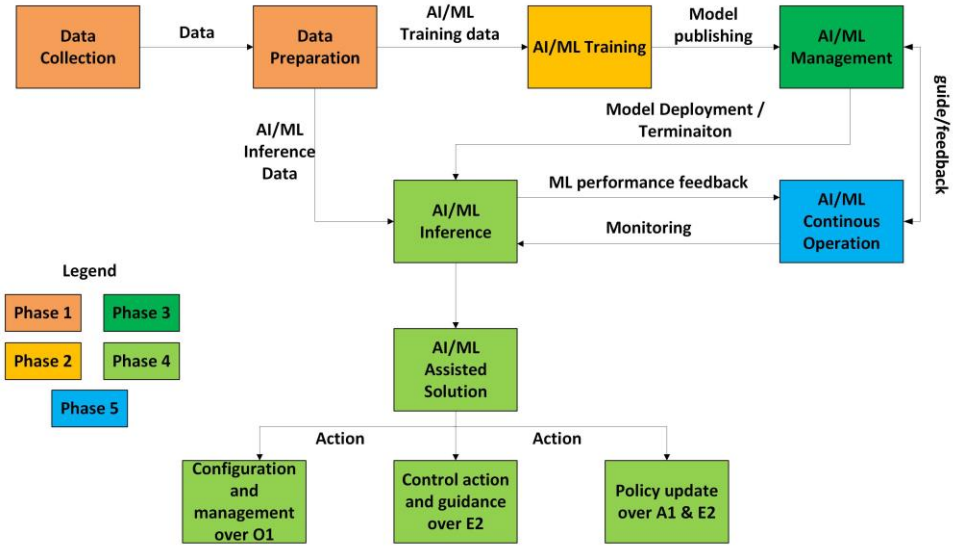


Figure 2.2. AI/ML lifecycle in O-RAN

The chapter concludes with a review of the methods used to evaluate the accuracy and reliability of the proposed models. The author analyzes various performance metrics and emphasizes the importance of model generalization to ensure their operability across different geographical regions and under diverse base station configurations. Conclusions are drawn regarding the necessity of hybrid approaches that combine expert knowledge of the physical layer with the flexibility of deep learning. This theoretical foundation serves as a logical transition to the subsequent chapters, where the developed models are applied to solve specific problems related to anomaly detection and Quality of Service (QoS) prediction.

11. Definition of QoS-critical services

The third chapter of the dissertation is dedicated to a detailed investigation and definition of modern multimedia and interactive services that impose the highest requirements on wireless network performance. The exposition begins with an overview of the fundamental paradigm shift in quality assessment, justifying the transition from purely technical indicators (QoS) to the comprehensive Quality of Experience (QoE). It is pointed out that traditional QoS indicators, while necessary for network monitoring, often fail to reflect the actual satisfaction of the end user, especially in highly dynamic services such as virtual reality and cloud gaming. In this context, the subjective factors and psychological aspects of perception that determine the final Mean Opinion Score (MOS) values are analyzed [52].

Special attention is given to the specifics of virtual reality and its requirements for ultra-low latency and high bandwidth. The concept of immersion and presence in the virtual environment is investigated, defining the critical latency thresholds above which the effect of cybersickness occurs. The various types of delays in the system are analyzed, including motion-to-photon latency, and how they affect the overall Quality of Experience in 360-degree video streaming. The dissertation

proposes a classification of influence factors, divided into system, human, and context factors, tracing their individual and combined impact on the end user [57].

The second major group of services discussed in the chapter encompasses online games and cloud gaming. It is emphasized that, unlike passive video content, interactivity here introduces new dependencies where even minor fluctuations in network parameters (jitter) can drastically degrade the gameplay [61]. The relationship between video content complexity, scene dynamics, and the required bitrate to maintain satisfactory quality at high resolutions is investigated. The exposition also covers Cellular Vehicle-to-Everything (C-V2X) communication scenarios, where reliability and transmission speed are of critical importance for traffic safety, imposing strict requirements on the availability and capacity of the radio channel [71].

The chapter concludes with an analysis of the mathematical models for mapping between QoS and QoE indicators. Non-linear dependency functions are also examined, such as the logarithmic Weber-Fechner law and the exponential IQX hypothesis, which serve as the foundation for the predictive models developed in the subsequent chapters. Conclusions are drawn regarding the need for adaptive management mechanisms that account for the specific profile of each service in real time. This analysis serves to define the input parameters and optimization goals used in the design of the intelligent O-RAN environment and the subsequent experiments for improving the Quality of Service [56].

12. Design and implementation of a 5G/LTE test network based on the O-RAN standard

Understanding the architectural concepts applied in leading large-scale 5G/LTE testbeds provides a valuable theoretical and technological framework for designing custom test networks. The fourth chapter of the dissertation analyzes two of the most widely used test platforms worldwide: COLOSSEUM [76] and POWDER [92]. They serve as two notable examples that vividly demonstrate how the fundamental principles of Commercial Off-The-Shelf (COTS) hardware, softwarization, open source, virtualization, and containerization are successfully integrated to achieve realistic, scalable, and fully programmable research environments. COLOSSEUM, as the world's largest wireless network emulator, acts as a high-fidelity digital twin of O-RAN, providing the capability for safe data generation and AI/ML model training through a complex channel emulation system and large-scale computing infrastructure. On the other hand, the POWDER platform represents a large-scale "living lab" in an urban environment, which balances bare-metal hardware access for fundamental research with abstraction for higher network layers through a rich open-source software ecosystem and a flexible optical transport network. Based precisely on this in-depth analysis and the extracted architectural best practices, the custom experimental O-RAN testbed was designed, justified, and implemented to serve as a foundation for validating the algorithms proposed in the dissertation.

Its design adheres to the core concepts and architectural principles of openness and vendor neutrality, with the network built entirely using open-source software and COTS hardware. The network supports operation across all standardized frequencies from Frequency Range 1 (FR1), with experiments primarily focused on Band 7 (2.6 GHz) for LTE and Band n78 (3.5 GHz) for 5G NR. The developed network provides a wide range of telecommunication services, demonstrating the capabilities of a hybrid 4G/5G environment. It supports packet-switched mobile broadband services with QoS differentiation, short message delivery via the SGs interface to 3G core network components, as well as a rich set of IMS-based services through an integrated open-source platform, including VoLTE and SMS over SIP. Regarding 5G, the network supports enhanced Mobile Broadband (eMBB) access both in the transitional Non-Standalone (NSA) mode, utilizing a 4G core infrastructure, and in full Standalone (SA) mode with an independent 5G core network.

a. Network architecture

The physical architecture of the test network is built around a standard desktop computer equipped with a 10th-generation Intel Core i9 processor, 64 GB of RAM, an NVIDIA GTX 1650 GPU, and the Ubuntu operating system. This node hosts all core network components, including the radio access network elements for 5G gNB and LTE eNB, the hybrid core network, and the IMS platform. A separate server is used to implement the intelligent management, configured with two virtual machines via the VMware vSphere hypervisor. The first virtual machine hosts the components of the Near-Real-Time RAN Intelligent Controller (Near-RT RIC), while the second is dedicated to the Service Management and Orchestration (SMO) functionality. The radio interface is implemented using two USRP B210 software-defined radios (SDR), connected to the host system via a USB 3.0 interface, providing a 2x2 MIMO configuration and supporting a wide frequency range. To validate functionality, various User Equipment (UE) devices are used, including smartphones, LTE USB dongles, and a specialized Quectel RM500Q-GL 5G module, which provides access to detailed telemetry.

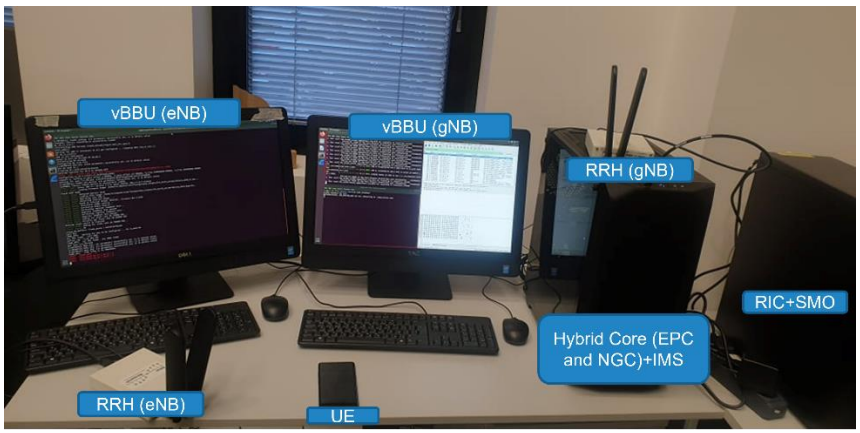


Figure 4.7. Physical architecture of the test network

The logical architecture represents a complex O-RAN implementation, integrating multiple software components in accordance with O-RAN Alliance and 3GPP specifications. The radio access network supports the simultaneous operation of LTE and 5G technologies, with the base stations capable of functioning both in a traditional monolithic mode and with a functional split according to Option 2 (at the PDCP/RLC boundary) and Option 7.2 (split between the low and high physical layers). The RAN is implemented using two alternative software platforms: srsRAN, which operates in a containerized environment via Docker, and OpenAirInterface (OAI), installed directly on the host machine's operating system (bare-metal configuration), allowing for optimal performance and direct access to hardware resources. The core network is based on Open5GS, integrating EPC and 5GC functionalities, and is complemented by a Kamailio IMS platform and Osmocom components to implement additional telecom services. All these core network components are containerized and consolidated within a common Docker Compose environment. Intelligent management in near-real-time (Near-RT RIC) is realized through three alternative implementations: FlexRAN (in a bare-metal configuration), FlexRIC (containerized via Docker), and the official reference implementation OSC Near-RT RIC, deployed as microservices on a Kubernetes cluster. Comprehensive orchestration is provided by the OSC SMO, which is also based on microservices and Kubernetes, integrating the Non-RT RIC components. Interaction among all these elements occurs via the standardized E2, X2, S1, N1/N2/N3, and O1 interfaces.

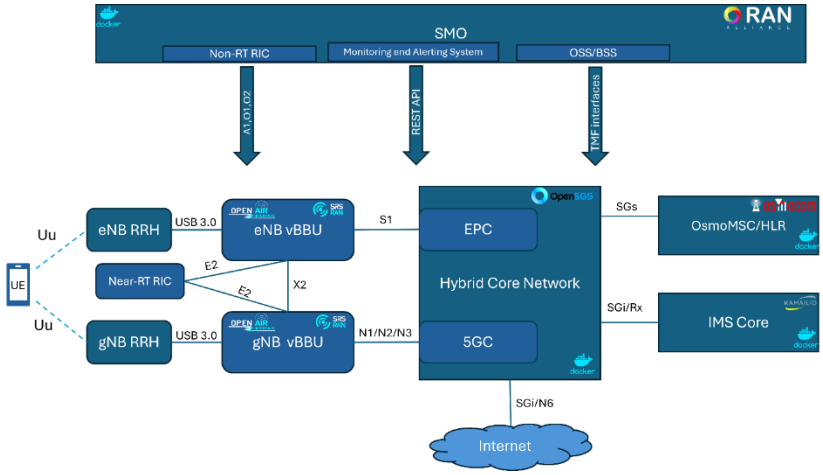


Figure 4.8. Logical architecture of the test network

b. Evaluation of network performance

The presented experiments were conducted without integrated intelligent management components, such as the Near-RT RIC and SMO, due to a lack of relevance to the baseline tests. Key Performance Indicators (KPIs) that are of paramount importance for Quality of Service-critical applications were selected for the evaluation. These indicators include the Round Trip Time (RTT), measured by sending ICMP packets to a Internet server (end-to-end) and within the radio access network (RAN); the downlink and uplink throughput using speed test tools; as well as the packet delay variation (jitter) over UDP connectivity.

i. Evaluation of LTE network performance

The experiments to establish the baseline performance indicators were conducted on the LTE configuration of the test network, using a Samsung Galaxy Note 9 smartphone as the User Equipment (UE). The measurements cover three base station architectures: a traditional monolithic eNB architecture without functional splitting, a functional split according to Option 2 (at the PDCP/RLC boundary), and a functional split according to Option 7.2 (split between the low and high physical layers). The network parameters include a 20 MHz bandwidth, corresponding to 100 Resource Blocks (RBs) in Frequency Division Duplex (FDD) mode. The operating frequencies are 2.67 GHz in the downlink with 64QAM modulation and 2.56 GHz in the uplink with QPSK modulation, with one connected user.

Parameter	Value
Bandwidth	20 MHz (100 PRBs)
Downlink frequency	2.67 GHz
Uplink frequency	2.56 GHz
Duplex method	FDD
Downlink modulation	64QAM
Uplink modulation	QPSK
Transmission mode	1
Number of connected users	1

Table 4.1. LTE network parameters

For the overall evaluation of the Round Trip Time (RTT), 1000 ICMP packets were sent to measure the end-to-end and RAN latency. The results reveal that the traditional monolithic eNB architecture consistently exhibits elevated latency levels, exceeding the values of the disaggregated Option 2 and Option 7.2 by 15-20%. The two functional split options demonstrate similar baseline characteristics, with RTT differences limited to 1-2 ms. However, for Option 7.2, 1.8% and 5.4% of packets with abnormal RTT values were registered during the end-to-end and RAN segment measurements, respectively. This frequency of deviations makes Option 2 more preferable for applications with hard real-time requirements, as it provides more reliable performance in the 99th percentile.

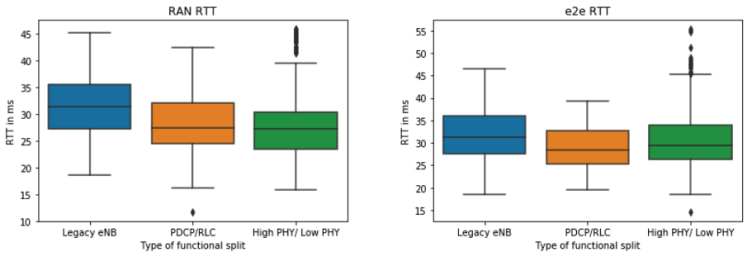


Figure 4.10. RTT across the different types of eNB architectures

The jitter evaluation, conducted via a 60-second UDP connection, shows that Option 7.2 achieves the lowest values in the downlink, reaching below 1 ms. A correlation is observed wherein centralizing more network functions (as in Option 7.2) leads to a narrower range of latency variations and more stable performance.

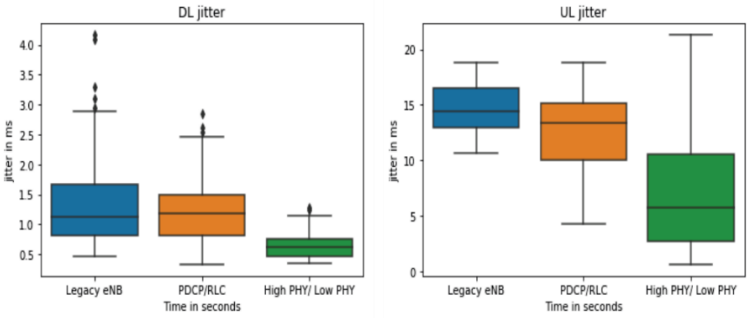


Figure 4.11. Jitter in downlink (DL) and uplink (UL) direction during a 60-second connection

To confirm this trend, long-term 60-minute measurements were also conducted. In these, the initial advantage of Option 7.2 is slightly maintained, but the differences among the three architectures become statistically insignificant over time.

This indicates that all three architectural approaches provide acceptable jitter values for the prolonged operation of jitter-sensitive applications.

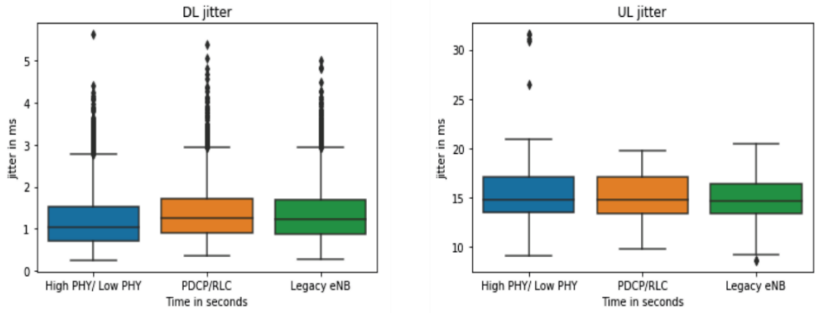


Figure 4.12. Jitter in downlink (DL) and uplink (UL) direction during a 60-minute connection

The throughput analysis demonstrates that the median speeds in the downlink (DL) are comparable across all configurations. In contrast, in the uplink (UL), a dramatic improvement of over 4 times is observed with the functional splits compared to the traditional monolithic architecture. This can be explained by the optimized resource management and more efficient scheduling in the centralized modules of the split architectures. The combination of significantly increased UL speed and lower latency makes disaggregated solutions highly suitable for modern symmetric services, such as holographic video conferencing and interactive systems, which require optimal and predictable communication.

	Legacy eNB				PDCP/RLC				High PHY/ Low PHY			
No.	Ping (ms)	Jitter (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	Jitter (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	Jitter (ms)	DL (Mbps)	UL (Mbps)
1	27	3	65.9	4.8	28	18	64.2	19.8	27	2	69.6	18.33
2	12	20	16.1	4.02	18	17	66.2	19.1	27	2	66.2	17.9
3	21	5	63	4.67	27	3	61.9	18.3	27	9	68.3	18.3
4	28	19	68.4	4.44	28	1	66.5	18.3	21	9	66.8	18.33
5	22	10	61.2	4.35	26	4	66.8	9.87	28	1	68.5	18.2
6	27	23	32.3	4.23	28	1	67.6	18.3	2	2	16.08	18.3
Average	22.83	13.33	51.15	4.42	25.83	7.33	65.53	17.28	26.33	4.17	59.25	18.23
Median	24.5	14.5	62.1	4.395	27.5	3.5	66.35	18.3	27	2	67.55	18.3

Table 4.2. Network throughput in LTE mode

c. Evaluation of 5G network performance in SA mode

Experiments were also conducted with the test network configured as a 5G SA (Standalone) network to evaluate the impact of different base station (gNB) architectural variants on Key Performance Indicators (KPIs). The same three architectures were used as in the LTE tests: traditional monolithic, Option 2, and Option 7.2. To ensure direct comparability, the network parameters are identical for all configurations and are based on 3GPP specifications. The configuration includes a 40 MHz bandwidth at a frequency of 3.5 GHz in Time Division Duplex (TDD) mode, using 64QAM modulation in the downlink and QPSK in the uplink with one connected user.

Parameters	Value
Bandwidth	40 MHz (106 resource blocks and 30 KHz SCS)
Downlink frequency	3.5 GHz
Uplink frequency	3.5 GHz
Duplex method	TDD
Downlink modulation	64QAM
Uplink modulation	QPSK
Transmission mode	1
Number of connected users	1

Table 4.3. 5G network parameters

The latency evaluation was again carried out through Round Trip Time (RTT) measurements at two levels: within the radio access network (RAN RTT) and end-to-end to an external internet server (e2e RTT). The obtained results show a slight advantage of the functional splits over the traditional monolithic architecture. For RAN RTT, both split configurations demonstrate lower values, with Option 7.2 achieving a reduction of approximately 6%, while the differences between Option 2 and Option 7.2 remain almost negligible. In the end-to-end measurements, Option 2 shows a slight advantage, while Option 7.2 reaches median values close to the traditional architecture, but with slightly higher deviations in the upper range. These variations can be explained by the specific mechanisms of the physical layer split or by performance limitations of the used hardware.

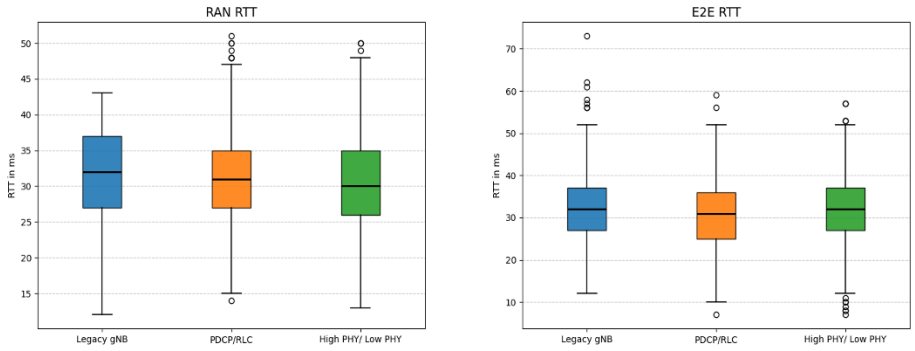


Figure 4.13. RTT across the different types of gNB architectures

Jitter measurements were conducted through long-term 60-minute UDP sessions for both transmission directions. In the downlink (DL), Option 7.2 demonstrates the lowest jitter values, which in some cases are nearly 1.5 times lower compared to the monolithic architecture. Option 2 also records a moderate improvement, proving that centralizing a larger volume of functionalities and the constant bitrate in the fronthaul lead to more stable latency in the downlink direction. In the uplink (UL), however, Option 7.2 registers significantly higher jitter values compared to the other architectures. This confirms that jitter is particularly sensitive to the distribution of physical layer functions and that a higher degree of centralization requires proper functioning to achieve stability.

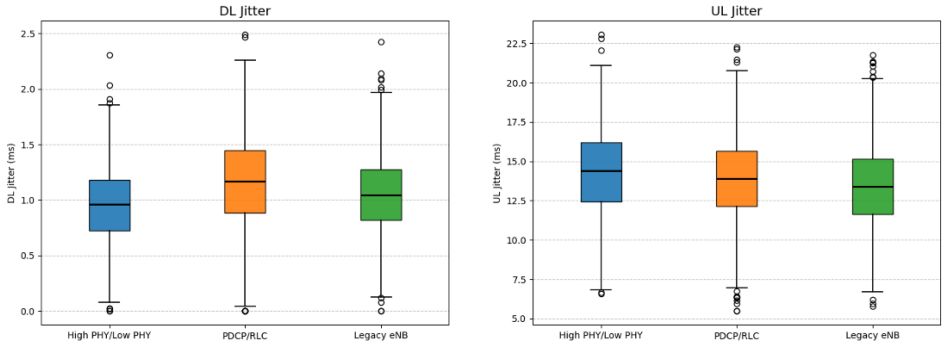


Figure 4.14. Jitter values in downlink and uplink directions across the different gNB architectures

Regarding the measured throughput, the differences among the three architectures proved to be insignificant, with all providing stable performance that, quite expectedly, exceeds the throughput achieved in LTE mode.

	Legacy gNB			PDCP/RLC			High PHY/ Low PHY		
No	Ping (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	DL (Mbps)	UL (Mbps)	Ping (ms)	DL (Mbps)	UL (Mbps)
1	20	81.95	36.65	20	86.05	38.48	24	84.33	37.71
2	20	77.22	36.59	20	81.13	38.42	22	81.27	37.57
3	20	85.03	36.53	22	86.73	37.27	21	83	36.52
4	20	80.34	36.13	20	78.73	36.5	23	77.94	36.87
5	20	82.01	36.53	24	81.97	37.27	20	82.77	37.64
6	20	83.61	36.43	20	84.28	37.16	20	81.13	36.79
Avg	20	81.69333	36.47667	21	83.14833	37.51667	21.66667	81.74	37.18333
Median	20	81.98	36.53	20	83.125	37.27	21.5	82.02	37.22

Table 4.15. Network throughput in 5G SA mode

13. Automated anomaly detection in mobile networks

The fifth chapter of the dissertation is dedicated to the design, implementation, and evaluation of intelligent anomaly detection systems, which are critical for ensuring the security and reliability of mobile networks. The theoretical introduction in the chapter justifies the need for automated approaches due to the increasing complexity of network architectures and the impossibility of manual monitoring of the massive volumes of generated data. The taxonomy of anomalies is examined, including point, contextual, and collective anomalies, emphasizing that in modern telecommunication systems, they are often indicators of cyberattacks, hardware failures, or software bugs. The transition from traditional statistical methods to models based on deep learning is defined as a mandatory prerequisite for the effective recognition of complex and previously unknown patterns of atypical behavior in network traffic. The chapter also presents work from three scientific publications that cover various aspects of the problem, from theoretical analysis and comparison of approaches to their practical implementation.

a. Development and implementation of algorithms for the timely detection of anomalies in mobile networks

First, a deep learning algorithm for anomaly detection is presented, which can be executed as an rApp application on the Non-RT RIC component of the O-RAN architecture. Utilizing radio access network-specific Key Performance Indicators (KPIs) obtained from the E2 nodes, the rApp monitors long-term trends and patterns in terms of performance and trains a supervised deep learning model based on them. Upon detecting abnormal network behavior or suboptimal performance, the application can take corrective actions by sending reconfiguration instructions over the A1 interface to the Near-RT RIC, which in turn controls the base station where the anomaly occurred.

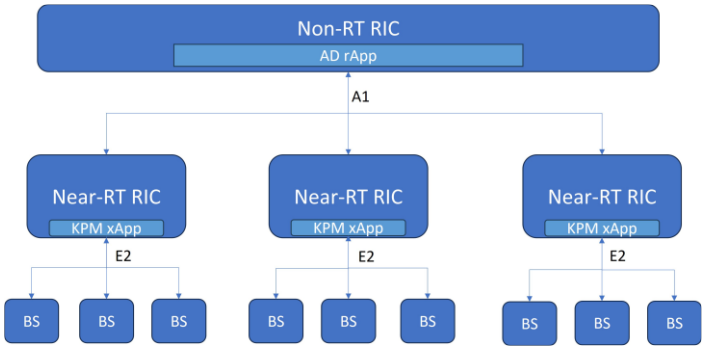


Figure 5.1. Schematic of an anomaly detection rApp

A prototype of the proposed solution was developed using Jupyter Notebook, Python 3.7.2, Keras, and TensorFlow. The performance evaluation was conducted in a typical scenario involving network monitoring via a KPM xApp running on the Near-RT RIC. The data used for training the model was collected from a real LTE network and consists of two-week logs from a group of 10 base stations in total, with a reporting interval of 15 minutes. Although the data originates from a traditional LTE architecture, the same parameters can be recorded during the monitoring cycle in O-RAN networks. Each sample in the dataset contains features such as a timestamp, unique cell identifier, percentage utilization of Physical Resource Blocks (PRBs) in the downlink and uplink, average and maximum transferred traffic (in Mbps), as well as the average and maximum number of simultaneously active User Equipment (UE) devices. For the purposes of supervised learning, a label is also included, where a value of zero indicates normal operation and a value of one indicates anomalous behavior. The final neural network architecture was selected after extensive hyperparameter fine-tuning.

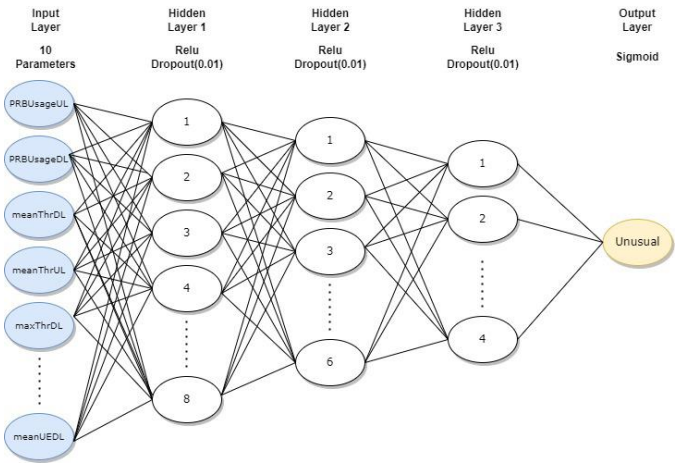


Figure 5.2. Architecture of the selected neural network

The architecture of the final model consists of an input layer that receives and processes the raw data (excluding the timestamp, the cell name, and the maximum number of active devices in total for both directions), followed by three hidden layers. These layers contain 8, 6, and 4 neurons, respectively, equipped with the computationally efficient Rectified Linear Unit (ReLU) activation function, which helps capture complex patterns and mitigates the vanishing gradient problem. To prevent overfitting, Dropout layers are added after each hidden layer, which randomly deactivate a portion of the neurons during training, thereby improving the model's generalization. The output layer consists of a single neuron with a sigmoid activation function, which converts the output into a probability score for binary classification, indicating the likelihood that the input sample represents an anomaly.

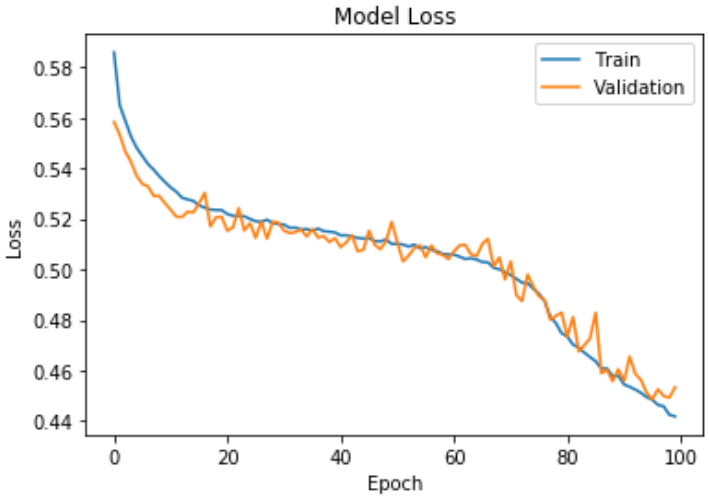


Figure 5.3. Model loss during the training and validation process

Prior to training, the dataset was divided into 70% for training, 10% for cross-validation, and 20% for testing. The model was trained for 100 epochs with a batch size of 32 examples and a learning rate of 0.01. An analysis of the loss graph shows that the training loss decreases with each epoch, confirming effective learning. The validation loss initially follows a similar trend, with minor deviations of about 1% observed, which do not significantly affect the model.

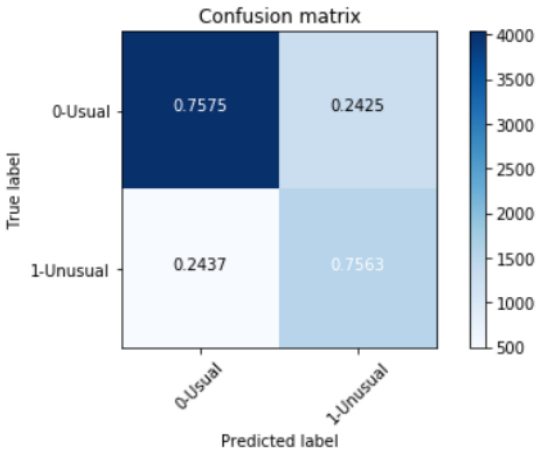


Figure 5.4. Confusion Matrix

Following performance evaluation, the algorithm achieves an overall accuracy of 0.7571, a precision of 54.43%, a recall of 0.7563, and an F1 score of 0.6630. Although the confusion matrix reveals a well-balanced ratio between false positives and false negatives, the model proves ineffective for practical purposes due to the high number of misclassifications. The low precision indicates that a large volume of normal traffic would be classified as anomalous, which would overload the network with unnecessary reconfiguration messages, while false negatives would lead to undiagnosed degradation of critical services.

Another significant drawback of the proposed approach is its reliance on supervised learning. To function effectively, this method requires pre-labeled datasets, which are extremely difficult to collect in dynamic network environments due to the inherent class imbalance, where normal data always dominates over rare anomalies. This limitation necessitates a shift toward semi-supervised or unsupervised learning approaches. These methods are trained solely on normal operational data and are capable of detecting both known and novel, atypical anomalies. Consequently, future research is focused on the development and application of unsupervised learning algorithms, which provide a higher degree of automation and reliability in the management of modern mobile networks.

b. Transformer-based reconstruction model for anomaly detection

In the second part of the chapter, a reconstruction-based model is presented, which utilizes the Transformer architecture for anomaly detection in the operational data of a mobile network. The development of the model is motivated by the proven success of Transformers in detecting anomalies in ECG data [124], where their ability to capture unusual patterns at both the local and global levels proves to be highly effective. For comparison of the results, the standard LSTM-Autoencoder architecture is used, which follows the same reconstruction principle. The architecture consists of an encoder with two LSTM layers that compress the input data into a narrow four-dimensional latent representation (bottleneck). After restoring the dimensionality via a RepeatVector, a decoder with two LSTM layers reconstructs the original signal, and an output dense (fully connected) layer realizes the final reconstruction. The model maintains short-term and long-term memory, allowing for efficient modeling of temporal dependencies. Additionally, the One-Class SVM algorithm, which constructs a separating hyperplane, as well as the statistical IQR and Z-score methods, were applied.

In training and testing the models, the telecommunications component of a large-scale dataset for the city of Milan and the province of Trentino was used. The geographical territory of these two regions is divided into grids corresponding to squares of approximately 235 x 235 meters, with Milan consisting of 1000 squares and Trentino of 6575. Telecom Italia's Semantic and Knowledge Innovation Lab provided the Call Detail Records (CDR) through which traffic values were measured. Every time a user interacts with the network, a new CDR entry is created, containing the time of interaction and the serving base station. The geographic location of the user is determined via coverage maps showing the territory served by each base station, with interactions aggregated according to the grid square to which they belong. The records are temporally aggregated into ten-minute time intervals and multiplied by an operator-defined constant to hide the true number of calls. The dataset provides geo-referenced measurements for a two-month period from November 1, 2013, to January 1, 2014. Features include the gridID, time interval, as well as the number of incoming and outgoing SMS messages, phone calls, and the level of internet traffic. The initial data consists of 62 text files, which were merged and aggregated so that each sample represents an hourly measurement. To reduce computational complexity in the early stages

of development, the dataset was restricted to 6 of the 10 base stations with the highest internet usage, and non-essential features were removed. Due to inconsistencies in the measurements after December 22, likely caused by the holiday season, these anomalous data points were excluded. The final dataset, consisting of 8928 records, was sorted chronologically and split: the training set includes data up to December 11, 2013, and the test set covers the remaining measurements up to December 22. To incorporate temporal information, features such as hour, day of the week, day of the month, and month were extracted, after which the data was standardized and normalized.

Anomaly type	GridID	Date	Time	Anomaly label
Internet spike	5059	14.12.2013	10am-8pm	1
SMSIn drop	All grids	18.12.2013	10am-8pm	2
CallOut drop	All grids	16.12.2013	10am-8pm	3

Table 5.1. Description of the artificially injected anomalies

Since the original dataset consists of normal traffic profiles, initially, all measurements were assigned an anomaly label of 0. To test the models, artificial anomalies simulating real network issues were injected into the test set. A spike in internet consumption simulating a DDoS attack was introduced, as well as sudden drops in incoming SMS messages and outgoing voice calls, simulating a network failure. These 143 anomalies were strategically introduced during peak hours between 10:00 and 20:00, when network traffic is at its highest.

The architecture of the proposed Transformer model utilizes only the encoder block from the original architecture. By training solely on normal data, the model learns their distribution and subsequently detects anomalies by calculating the loss function between the reconstructed and the original data. The architecture consists of a two-layer encoder, where the first layer creates hidden representations that are fed to the second layer to build higher-level representations. Each transformer block includes a multi-head self-attention mechanism and a position-wise feed-forward neural network, equipped with residual connections and layer normalization. The detection process involves training on clean time series, defining a threshold for the reconstruction error, and flagging any deviation above this threshold as an anomaly when testing with new data.

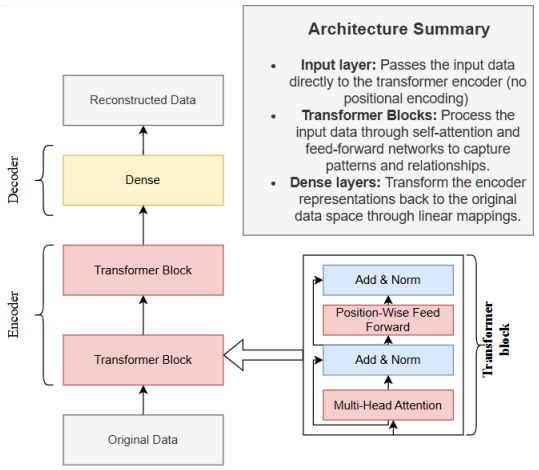


Figure 5.5. Transformer architecture

In evaluating the results, the Transformer demonstrates exceptional reconstruction performance. After training for an optimal 23 epochs to avoid overfitting, the model successfully detects 138 of the 143 injected anomalies, achieving an accuracy of 96.5% with an execution time of just 29.45 seconds. In comparison, the LSTM-Autoencoder detects 133 anomalies, achieving 93.01% accuracy, but requires 40 training epochs and significantly more time (77.13 seconds) to overcome overfitting. The One-Class SVM algorithm performs surprisingly well, identifying 123 anomalies (86.01% accuracy) without the need for complex training. On the other hand, the statistical IQR and Z-score methods prove completely ineffective, detecting only 10 anomalies (6.99% accuracy), predominantly in internet traffic due to their significant deviation from median values.

Модел	Брой епохи	Брой открити аномалии	Точност	Време за изпълнение
Трансформатор	23	138	96.50%	29.45s
LSTM-Autoencoder	40	133	93.01%	77.13s
OC-SVM	-	123	86.01%	0.266s
Z-Score	-	10	6.99%	0.215s
IQR	-	10	6.99%	0.111s

Table 5.2. Summary of results

c. Implementation of an ML-based anomaly detection model in the O-RAN test network

At the end of the chapter, the practical implementation of the developed Transformer model for anomaly detection in the constructed experimental O-RAN environment is discussed. The algorithm is integrated as an analytical component into the 5G core network through the Network Data Analytics Function (NWDAF), which provides optimization services according to 3GPP specifications (Release 17). The microservice implementation of NWDAF from OpenAirInterface is used, which is logically divided into three layers: an Exposure layer for communication with external clients and event notification, a Monitoring layer for collecting telemetry from the core network (Open5GS) and the radio access network (via an xApp and the near-RT RIC), and an Analytics layer, where the Transformer model is executed as a standalone microservice. The workflow involves the continuous extraction of traffic data, which is analyzed in real time. An anomaly is registered when the calculated reconstruction error exceeds a predefined threshold, after which the system automatically generates a standardized notification for abnormal behavior. The practical verification proves the high reliability of the constructed architecture.

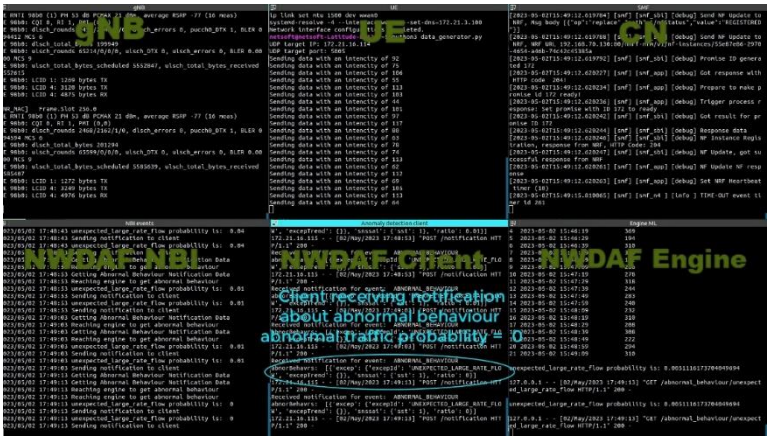


Figure 5.10. Visualization of normal internet traffic. NWDAF reports no anomalies.

In a normal traffic scenario, the NWDAF correctly reports an absence of deviations with a minimal reconstruction error. Upon simulating an atypical load via artificially injected traffic, the model reacts instantaneously, with the error values rising sharply, and in just a few seconds, the calculated probability of an anomaly reaches 97%. These results categorically confirm the ability of the integrated ML model to track network dynamics in a real 5G environment and to classify the escalation of anomalies with exceptional precision.

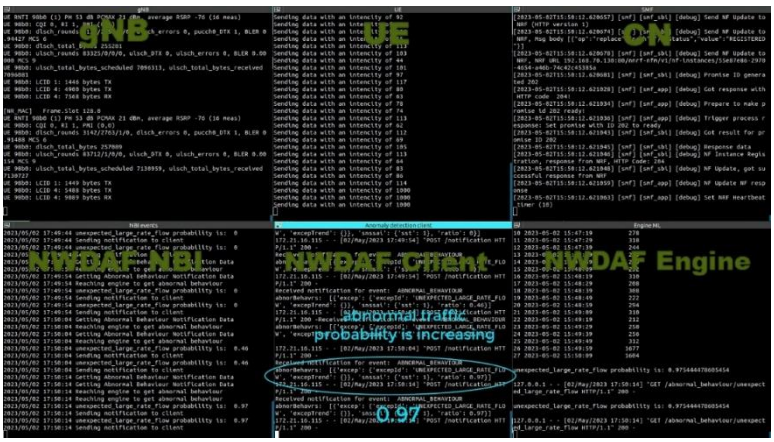


Figure 5.12. Classification of traffic as anomalous. NWDAF reports a 97% probability of abnormal behavior.

14. Models for the evaluation and prediction of QoE in interactive multimedia services

The sixth chapter of the dissertation examines the fundamental need for a transition from a static to a dynamic and predictive evaluation of the Quality of Experience (QoE) in modern mobile networks. The theoretical introduction justifies that with the rapid development of interactive multimedia services such as cloud gaming and virtual reality (VR), conventional reactive mechanisms for network optimization are no longer sufficient. Unlike traditional video, these services are extremely sensitive to the dynamics of the radio channel and require a combination of ultra-low latency, minimal jitter, and zero packet loss. Even a millisecond degradation can lead to physical discomfort for VR users or a loss of control in gaming. Therefore, resource management in the Open RAN architecture must be proactive; the network must not merely monitor the current state but predict the future behavior of the service based on historical data and current telemetry. In this context, machine learning algorithms are establishing themselves as the only reliable tool for modeling the complex non-linear and temporal dependencies between QoS and QoE, overcoming the limitations of classical analytical mathematical models.

To provide empirical data for the training and validation of the predictive models, a specialized system for continuous monitoring was designed and integrated into the O-RAN test network. The system is implemented using the Prometheus, Telegraf, and Grafana software stack and is entirely user-centric, as it extracts telemetry directly from the end devices. Two large-scale datasets were collected in a real mobile environment over a period of eight weeks, containing over 80,000 time samples each. The first dataset covers the transmission of 4K UHD VR 360-degree videos, recording throughput, latency, and packet loss. The second dataset is focused on the streaming of 2K gaming video at 60 FPS with high scene dynamics, additionally recording jitter values.

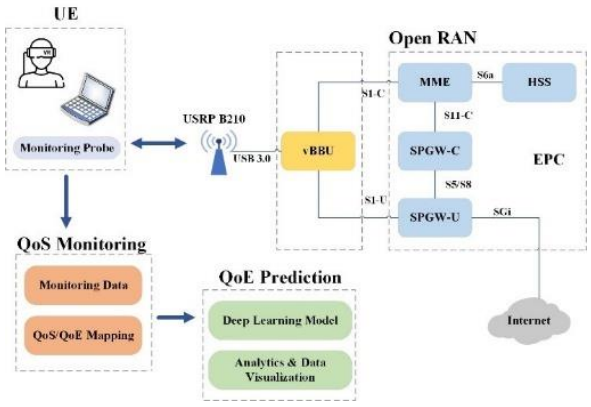


Figure 6.1. System for user-centric monitoring of the Quality of Service (QoS)

a. QoE prediction model for gaming video streaming

One of the scientific contributions in this chapter is the development of an innovative QoE prediction model for gaming video streaming, based on a Multi-headed Convolutional Neural Network (Multi-headed CNN) architecture. The model is specifically designed for processing multivariate time series by dividing the input sequences into parallel pathways (heads). Each pathway processes a 1D time series of a specific QoS parameter (latency, jitter, packet loss, or throughput). This approach allows the model to capture specific local anomalies, such as sudden spikes in latency, which would merge and be lost when using a standard CNN with a single common input. After filtering in the parallel pathways, the extracted features are concatenated into a single vector and fed into dense layers, which model the global dependencies and generate the final predicted Mean Opinion Score (MOS) for the perceived quality.

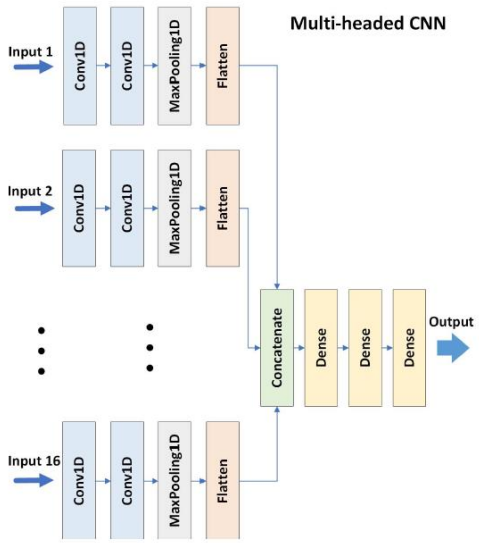


Figure 6.5. Architecture of the Multi-headed CNN

A critical step prior to feeding the data into the neural network is the preliminary transformation of raw QoS metrics into comparable QoE values. Since the perceived quality does not change linearly with respect to network parameters but rather demonstrates distinct threshold effects, non-linear mapping functions are integrated into the model. The use of logistic curves and the exponential IQX hypothesis (in combination with the psychophysical laws of Fechner and Stevens) provides precise mathematical mapping, which facilitates the training process of the neural network. The model utilizes a historical window of 48 time steps (two days) to predict the Mean Opinion Score for the next 24 time steps.

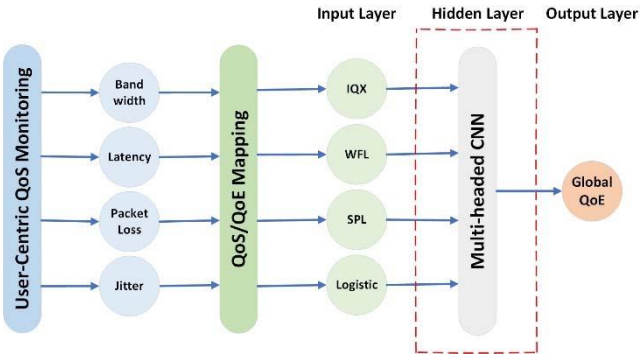


Figure 6.6. QoS to QoE (MOS) mapping

To evaluate the effectiveness of the developed model, an in-depth comparative analysis was conducted against other established deep learning architectures, including standard CNN, LSTM, Bidirectional-LSTM, and hybrid ResCNN-LSTM networks. The empirical results categorically prove the superiority of the proposed Multi-headed CNN architecture. The model achieves the lowest error rates across all key statistical metrics, registering a Mean Absolute Error (MAE) of 0.09786 and a Mean Absolute Percentage Error (MAPE) of just 2.52%, whereas competing LSTM and standard CNN models show errors in the range of 3.8% – 4.0%. These results confirm that the parallel convolutional structure is an extremely reliable tool for QoE prediction in real O-RAN networks, as it successfully combines local sensitivity to short-term network fluctuations with global generalization capability, making the model ideal for integration into near-real-time intelligent radio resource management controllers.

	MSE	RMSE	MAE	MAPE (%)	MedAE
CNN	0.03241	0.18005	0.14492	3.92163	0.13820
Multi-channel CNN	0.03432	0.18527	0.14835	4.03715	0.13218
Multi-headed CNN	**0.01342**	**0.11588**	**0.09786**	**2.52839**	**0.09664**
TCN	0.03581	0.18924	0.15303	4.14064	0.13052
LSTM	0.03250	0.18029	0.14196	3.85048	0.13774
ResCNN-LSTM	0.03245	0.18016	0.14317	3.87785	0.13492
RNN	0.03450	0.18576	0.14212	3.88562	0.11938
GRU	0.03232	0.17978	0.14060	3.82902	0.12912

Table 6.2. Evaluation of the Quality of Experience (QoE) prediction model

b. QoE prediction model for VR 360-degree video

Predicting the Quality of Experience (QoE) for VR 360-degree video presents a significantly more complex challenge compared to gaming video streaming, as virtual reality imposes even more extreme requirements on the network. To ensure a stable immersive experience, applications require stable and high throughput, minimal end-to-end latency, low jitter, and near-zero packet loss. Virtual reality is highly sensitive to momentary anomalies that can cause degradation in visual quality, delayed reactions, reduced resolution, or physical discomfort (cybersickness), which positions it among the most critical multimedia services. Due to this complexity, classical methods for static QoS/QoE mapping prove completely insufficient, as they cannot describe the cumulative temporal effects on perception. For precise prediction in this context, an innovative model based on an LSTM encoder-decoder architecture has been developed, which specializes in processing long time sequences and capturing contextual dependencies within the structure of network degradation.

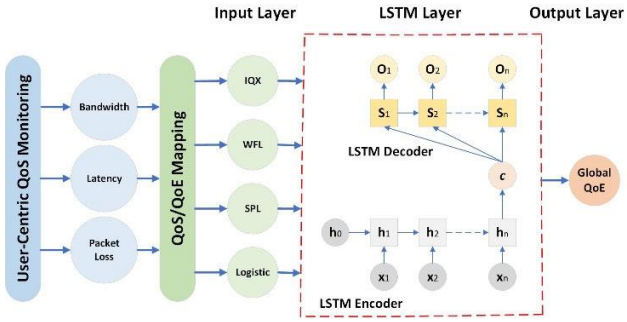


Figure 6.10. Architecture of the deep learning Quality of Experience prediction model

The proposed model operates through two interconnected components: an encoder and a decoder. The encoder takes as input a sequence of QoS parameters measured over the preceding 48 time steps and processes them through layers of Long Short-Term Memory (LSTM) cells. The primary objective of this component is to compress the historical information into a compact and rich latent space that reflects the dynamic structure of temporal dependencies, including past trends, smooth transitions, and signs of potential network congestion. The decoder, in turn, utilizes this latent representation to extrapolate the dependencies and generate a prediction for future QoE values with a horizon of up to 24 steps ahead (24 hours). This allows intelligent network controllers to react proactively before actual quality degradation occurs. The input data to the model is generated through a preliminary non-linear transformation of the QoS parameters into their respective QoE equivalents, with the latent representation serving as a kind of "memory" to track the overall temporal structure of the traffic.

Unlike the Multi-headed CNN architecture, which was successfully applied in gaming scenarios, the LSTM model focuses on long-term dependencies. While convolutional networks are highly effective at isolating local variations, VR content requires the recognition of fine temporal structures spanning multiple samples. For example, a gradually increasing jitter or slightly elevated packet loss that foreshadows a severe crash. The architecture achieves this through the "remember" and "forget" mechanisms built into the gates of the LSTM cells, allowing the neural network to dynamically determine the weight of different historical moments.

	MSE	RMSE	MAE	MAPE (%)	MedAE
Naive	0.29636	0.54439	0.49610	15.2277	0.46343
Simple RNN	0.04126	0.20312	0.17529	4.68265	0.17242
LSTM	0.03383	0.18395	0.15186	4.08651	0.13727
Autoencoder LSTM	0.03134	0.17704	0.14987	4.05443	0.14518
Bidirectional LSTM	0.029015	0.17033	0.13958	3.81615	0.12350
Encoder-Decoder LSTM	**0.02541**	**0.15943**	**0.12491**	**3.42698**	**0.09981**
GRU	0.02767	0.16361	0.13198	3.63324	0.10320

Table 6.3. Evaluation of the Quality of Experience (QoE) prediction model

The training process is implemented by optimizing the Mean Squared Error (MSE) on the large-scale dataset of real data collected from the VR tests in the O-RAN environment. Empirical experiments prove that the optimal balance between precision and generalization is achieved when using exactly two LSTM layers, with dropout layers and early stopping mechanisms integrated to prevent overfitting. The results of the comparative analysis categorically emphasize the advantages of the developed architecture. The LSTM encoder-decoder model not only successfully follows the overall QoE dynamics up to 24 steps ahead but also demonstrates the highest accuracy among all tested state-of-the-art methods. The model achieves a Mean Squared Error (MSE) of 0.02541, a Mean Absolute Error (MAE) of 0.12491, and a Mean Absolute Percentage Error (MAPE) of just 3.42%. These values significantly outperform competing architectures such as Simple RNN, standard single-layer LSTM networks (MAPE 4.08%), Autoencoder LSTM, and Bidirectional LSTM models. The results prove that the proposed LSTM encoder-decoder model is a highly effective and reliable tool for proactive resource management for critical VR services in future intelligent mobile networks.

15. QoS prediction in C-V2X scenarios via machine learning

The seventh chapter of the dissertation examines the problem of Quality of Service (QoS) prediction in Cellular Vehicle-to-Everything (C-V2X) communication scenarios. The theoretical introduction justifies the critical role of C-V2X technologies for autonomous driving, where Ultra-Reliable Low-Latency Communication (URLLC) is a matter of functional safety, not merely user comfort. In this context, the network's ability to provide Predictive QoS is of key importance, allowing autonomous systems to proactively adapt their control strategies before actual communication degradation occurs. Despite the potential of machine learning, existing research often lacks realism, relies excessively on GPS coordinates (leading to the "memorization" of local topology), and fails to demonstrate the ability to generalize across different Mobile Network Operators (MNOs). To overcome these limitations, the chapter focuses on the development and experimental validation of universal, operator-agnostic models that predict downlink and uplink throughput relying solely on mobile network parameters available at the User Equipment (UE) level.

The large-scale "Berlin V2X Dataset," collected in a real operational urban environment, was used for the training and validation of the proposed algorithms. The measurements cover diverse scenarios (residential areas, parks, highways, and tunnels) and two separate MNO networks, allowing the capture of the full complexity of interference, fading, and dynamic cell load. The data correlation analysis reveals fundamental differences in the operators' radio resource management strategies. One operator demonstrates aggressive spectrum utilization with high variability, while the second applies a more conservative policy. The analysis isolates the key predictive features, establishing that downlink speed correlates strongly with the physical parameters of the signal, most notably the Signal-to-Noise Ratio (SNR), whereas the uplink connection is predominantly determined by signal strength indicators (RSRP, RSSI). At the same time, it is proven that delay barely correlates with radio parameters, confirming that latency depends primarily on the load of the utilized infrastructure.

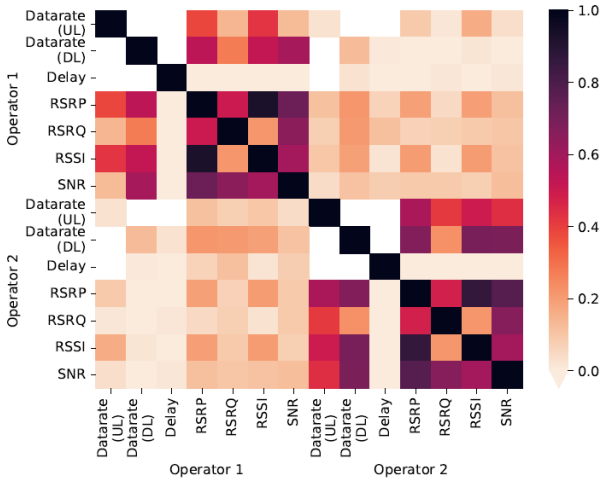


Figure 7.3. Correlation matrix

A critical architectural decision in the methodology of the proposed models is the complete exclusion of geographical coordinates from the input training vector. The feature engineering is based entirely on combining information from the physical layer (RSRP, RSQ, SNR) and the Medium Access Control (MAC) layer (Transport Block Size and Modulation and Coding Scheme). This approach ensures that the algorithms learn the actual physical dependencies of radio propagation and the behavior of the network scheduler, enabling them to function independently of location and to generalize successfully to new network infrastructures. The feature importance analysis confirms that Transport Block Size, network latency, and jitter play the most significant role in prediction accuracy, indicating that the logical load of the cell is more critical than pure physical parameters.

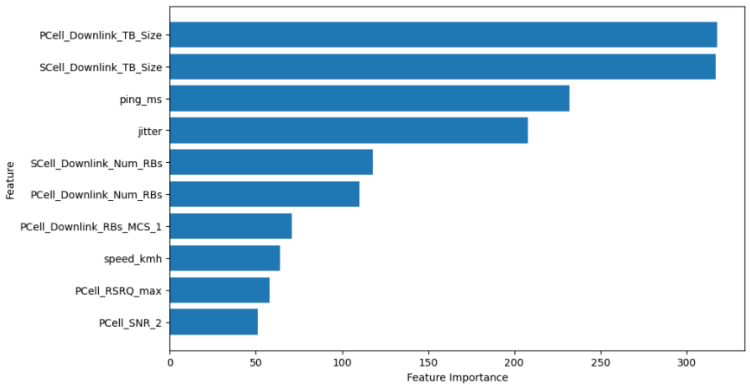


Figure 7.5. Top 10 most important parameters

The experimental phase compares four different classes of algorithms (Linear Regression, Random Forest, LightGBM, and XGBoost), with the models trained on data from one operator and validated directly on the data of the second. The results in predicting downlink throughput categorically prove the superiority of tree-based ensemble methods over linear approaches. The LightGBM algorithm demonstrates overall performance with a Coefficient of Determination (R^2) of 0.935 and a Mean Absolute Error (MAE) of 3.11 Mbps on the test data, offering an optimal balance between accuracy and computational efficiency. The Random Forest model performs almost identically (R^2 of 0.934 and MAE of 2.98 Mbps), whereas linear regression suffers a complete failure with a negative R^2 and a massive error of 22.4 Mbps, which proves the highly non-linear nature of the processes in the downlink channel.

Модел	R^2(Train)	R^2(Test)	MAE (Train) [Mbps]	MAE (Test) [Mbps]
LightGBM	0.967	**0.935**	4.29	3.11
Random Forest	**0.994**	0.934	**1.70**	**2.98**
XGBoost	0.980	0.924	3.52	3.31
Linear Regression	0.816	-773.38	12.70	22.40

Table 7.1. Results of downlink throughput prediction

In predicting the uplink throughput, the network dynamics change, due to the limited power budget of the User Equipment (UE). In this scenario, Random Forest emerges as the most accurate model, achieving the lowest error of 1.59 Mbps with an R^2 of 0.919, as its approach handles the specific noise in the uplink data more successfully. LightGBM achieves an identical R^2 but with a slightly higher error, while linear regression once again proves completely inapplicable. The conducted experiments categorically prove that the exclusion of GPS coordinates and the reliance solely on PHY/MAC radio parameters allow non-linear ensemble models to achieve over 90% accuracy even when tested on the network of a completely different operator. This successfully fulfills the main objective of the research - the creation of a reliable, operator-agnostic predictive QoS model, which is critically essential for the safe deployment of autonomous mobility in future communication networks.

Модел	R^2(Train)	R^2(Test)	MAE (Train) [Mbps]	MAE (Test) [Mbps]
LightGBM	0.998	0.919	0.878	1.68
Random Forest	**0.998**	**0.919**	**0.33**	**1.59**
XGBoost	0.99	0.88	0.77	2.49
Linear Regression	0.38	-1.73	9.64	14.53

Table 7.2. Results of uplink throughput prediction

16. Оптимизация на мрежовите ресурси и управление на качеството чрез мрежово нарязване

The eighth chapter of the dissertation examines the practical closing of the intelligent management loop in mobile networks, transitioning from the monitoring (Chapter 5) and prediction (Chapters 6 and 7) phases to the stage of actual optimization and execution. The exponential growth of network traffic necessitates the development of flexible architectures capable of managing dynamic resource sharing under strict access control. In this context, the concept of Network Slicing, introduced by the NGMN alliance, provides a fundamental mechanism for creating multiple independent logical networks (slices) over a shared physical infrastructure. Each slice is isolated and optimized for specific Quality of Service (QoS) requirements, allowing the coexistence of heterogeneous services. However, managing this architecture requires finding a complex balance between service customization, resource management efficiency, and overall system complexity. To overcome these challenges, the research focuses on two main directions: automated network slice selection through machine learning and dynamic radio resource optimization in a real O-RAN environment.

To realize automated network slice selection, a methodology for intelligent traffic classification has been developed, which proactively routes user requests to the appropriate network slice (eMBB, URLLC, or mMTC). The specialized "Deep Slice" dataset is used, containing key indicators extracted from control messages between the device and the network, such as delay tolerance, maximum packet loss, and supported technology. After preprocessing to remove redundant features, an analysis using a Decision Tree was conducted, which isolated the most important classifiers. It was established that a packet loss rate below 0.001 effectively distinguishes critical URLLC traffic, the supported technology separates eMBB from mMTC, and the delay budget refines the boundary between mMTC and URLLC. Initial training demonstrated 100% accuracy, which, however, is an indicator of severe overfitting due to the limited number of unique records in the raw dataset, which would compromise generalization in a real-world environment.

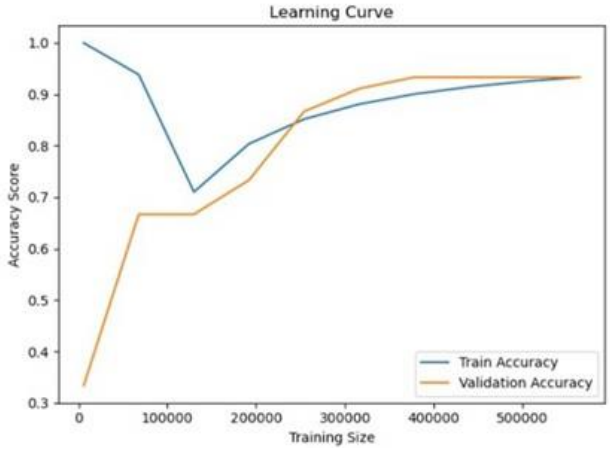


Figure 8.6. Learning curve after oversampling and deep feature engineering

To overcome the generalization problem and create a robust model, a process of deep feature engineering was applied. All input features were converted into a binary format, which initially lowered the accuracy to 91% but drastically improved the parallelism between the training and validation curves, ensuring adequate performance with unseen data. An additional challenge was the severe class imbalance, where URLLC traffic dominated nearly half of the records. This problem was resolved by applying an oversampling technique, which balanced the classes and increased the accuracy to 92%. The AdaBoostClassifier ensemble algorithm was selected as the final classifier, which uses sequentially connected decision trees to minimize errors. The model is extremely fast and computationally lightweight, making it ideal for implementation as an xApp application at the network edge. After training on 80% of the data, the final model achieves an accuracy of 93.4%, with the confusion matrix demonstrating a minimal and statistically insignificant number of false-positive errors, confirming the successful traffic classification.

	URRLC	eMBB	mMTC
Accuracy	0,934		
Precission	1	1	0,834
Recall	0.801	1	1
F-Score	0.89	1	0,91

Table 8.2. Key performance indicators of the model

The second direction of the research builds upon the successful classification through a practical demonstration of dynamic Physical Resource Block (PRB) allocation in a real O-RAN test network to guarantee QoS for multimedia services. The experimental setup utilizes the open-source FlexRAN platform in the role of the Near-RT RIC and integrates a specialized monitoring system for DASH video streaming.

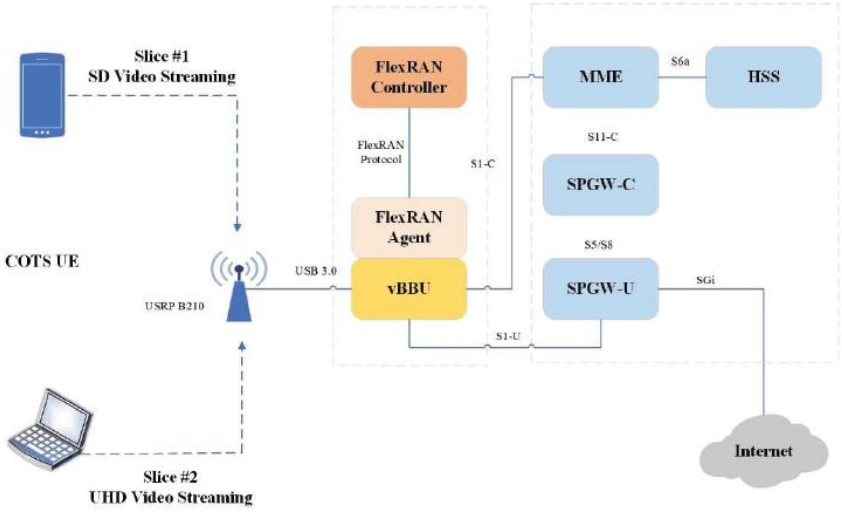


Figure 8.8. Experimental environment for Network Slicing

Two separate slices were created - one for Standard Definition (SD) video and one for Ultra-High Definition (UHD) video. With a static equal distribution of resource blocks (50% for SD and 50% for UHD), the results show that the resources are entirely sufficient for the SD stream (zero dropped frames and low latency) but are absolutely insufficient for the UHD video, which suffers critical degradation with unacceptable delays and 6897 dropped frames.

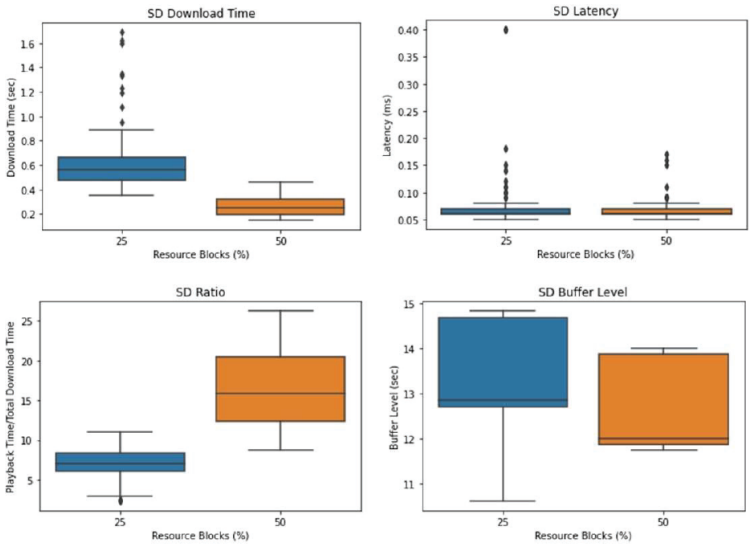


Figure 8.9. SD video KPIs under different radio resource allocations

To resolve this problem, an autonomous script was developed that continuously monitors the key indicators and, upon detecting degradation in the UHD stream, generates a trigger to the FlexRAN API. The controller reacts in real time by dynamically reconfiguring the network - it restricts the resources for the SD slice to 25% and preferentially expands the capacity of the UHD slice to 75%. The empirical results of this dynamic optimization are definitive: the reduction of resources for the SD video has a minimal, almost imperceptible impact (only 7 dropped frames), while the quality of the UHD streaming improves drastically. Download time and latency drop significantly, the buffer level stabilizes, and the number of dropped frames decreases by more than 43 times (down to 158). This experiment categorically proves that the integration of intelligent monitoring with dynamic network slicing at the RAN edge is a key mechanism for effective radio resource management and guaranteeing high quality for critical multimedia services.

SD (25% PRBs)	SD (50% PRBs)	UHD (50% PRBs)	UHD (75 % PRBs)
7	0	6897	158

Table 8.3. Dropped frames in SD and UHD video streaming

Conclusions:

The dissertation investigates and validates the hypothesis that integrating Artificial Intelligence (AI) and Machine Learning (ML) into the Open RAN architecture is not only possible but also necessary for the effective management of modern and future mobile networks. Through theoretical analysis, the design of a complex experimental environment, and the execution of a series of practical experiments, the dissertation demonstrates that intelligent algorithms can successfully solve the problems of complexity, dynamics, and heterogeneous service requirements that traditional management methods fail to address. Within the framework of the research, the following key contributions have been achieved:

5. **Creation of a realistic experimental platform:** A fully functional O-RAN-based test network integrating open-source components and Commercial Off-The-Shelf (COTS) hardware was successfully designed and deployed. This platform serves as a foundation for generating unique real-world datasets and for validating the developed models in an environment that reflects the actual hardware and software limitations of the network. The comparative performance analysis in LTE and 5G modes confirms that architectures with functional splits offer comparable performance to traditional monolithic solutions, but with added flexibility.

6. **Enhancing security and reliability through AI/ML:** The dissertation proved the superiority of deep learning models over classical statistical methods in detecting network anomalies. The developed Transformer model, integrated into the OAI NWDAF, demonstrated the ability to detect complex traffic anomalies with an accuracy of 96.5%, significantly exceeding the results of methods such as Z-score and IQR. This enables a proactive reaction in near-real time.
7. **Predictive QoE management for interactive services**: Specialized ML models for predicting the Quality of Experience (QoE) for services highly sensitive to network parameters were developed:
    - For **gaming video streaming**, a Multi-headed CNN model was validated, which effectively captures short-term fluctuations in jitter and latency, achieving the lowest prediction error.
    - For **VR 360-degree video**, an LSTM Encoder-Decoder model was created, capable of modeling long-term temporal dependencies and predicting quality up to 24 steps ahead.
    - For **C-V2X scenarios**, it was proven that location-agnostic models (such as LightGBM and Random Forest), trained solely on radio parameters without GPS coordinates, can predict throughput with high accuracy ($R^2 > 0.9$) even when transferred across networks of different mobile operators.
8. **Closing the management loop through automation:** The dissertation demonstrated the practical implementation of closed-loop control in an O-RAN environment. The developed mechanism for dynamic network slicing and resource reallocation, based on real-time monitoring, demonstrated the ability to restore the quality of a UHD video stream through automated traffic prioritization. Additionally, an ML model for the automatic classification and association of users to the correct network slice was validated with an accuracy of over 93%.

The results of the dissertation confirm that the Open RAN architecture provides the necessary tools and interfaces to transform the mobile network from a static infrastructure into an intelligent, adaptive platform. By utilizing deep learning models for prediction and detection, combined with the programmability of the RIC controllers, it is possible to achieve a level of optimization and service quality guarantee (QoS/QoE) that is unattainable with traditional methods. The developed models and algorithms are operator-agnostic and applicable in real-world scenarios, offering solutions to critical problems such as traffic management for VR and autonomous vehicles. This work contributes to both the theoretical understanding of AI-native networks and their practical realization by providing validated architectural solutions and software implementations.

References

[1] G. Heine and H. Sagkob, GPRS: Gateway to Third-Generation Mobile Networks. London, U.K.: Artech House, 2003.

[2] Colonna, Massimo & Barbaresi, Andrea & Zarba, Giovanna & Mantovani, Andrea. (2008). A Brief Survey of VoIP QoS over a multi-RAT Heterogeneous Wireless Network.

[4] Technical Specifications and Technical Reports for UTRAN-Based 3GPP System, 3GPP TR 21.10. 2003

[5] 3GPP, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Release 8 Overview,” Release 8, Dec. 2008. [Online]. Available: https://www.3gpp.org

[8] 5G NR Network Interfaces: Xn, NG, E1, F1, F2 Explained [Available at: https://www.rfwireless-world.com/tutorials/5g/5g-nr-network-interfaces] [Last Accessed: 22.05.2025]

[17] Study on new radio access technology Radio access architecture and interfaces, 3GPP, 3rd Generation Partnership Project, 3 2017, v14.0.

[21] Xu, F., Yao, H., Zhao, C. et al. Towards next generation software-defined radio access network–architecture, deployment, and use case. J Wireless Com Network 2016, 264 (2016). https://doi.org/10.1186/s13638-016-0762-6

[30] M. Polese, L. Bonati, S. D’Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," in IEEE Communications Surveys & Tutorials, vol. 25, no. 2, pp. 1376-1411, Secondquarter 2023, doi: 10.1109/COMST.2023.3239220.

[35] S. Suthaharan, "Supervised learning algorithms," in Machine Learning Models and Algorithms for Big Data Classification, New York, NY, USA: Springer, 2016, p. 183–206.

[36] Tyagi, Kanishka & Rane, Chinmay & Sriram, Raghavendra & Manry, Michael. (2022). Unsupervised learning. 10.1016/B978-0-12-824054-0.00012-5.

[37] A. Paz and S. Moran, "Non deterministic polynomial optimization problems and their approximations," Theoretical Computer, vol. 15, no. 3, p. 251–277, 1981.

[42] O-RAN AI/ML Workflow Architecture and Framework - ERK, https://erk.fe.uni-lj.si/2024/papers/cop(o_ran_ai_ml).pdf, [Last Accessed: 05.07.2025]

[52] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis and P. I. Lazaridis, "A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction," in IEEE Access, vol. 10, pp. 19507-19538, 2022, doi: 10.1109/ACCESS.2022.3149592.

[57] Canadian Radio-television and Telecommunications Commission (CRTC), “Telecom Decision CRTC 2018-241: CISC Network Working Group – Non-consensus report on quality of service metrics to define high-quality fixed broadband Internet access service,” Ottawa, Canada, 13 July 2018. [Online]. Последно достъпено на 13.07.2025 Линк: https://crtc.gc.ca/eng/archive/2018/2018-241.htm

[61]https://gamerhub.co.uk/gaming-industry-dominates-as-the-highest-grossing-entertainment-industry/ Достъпен на: 17.07.2026

[71] Cross-layer latency analysis for 5G NR in V2X communicationsHorta J, Siller M, Villarreal-Reyes S (2025) Cross-layer latency analysis for 5G NR in V2X communications. PLOS ONE 20(1): e0313772. https://doi.org/10.1371/journal.pone.0313772

[56] Alreshoodi, Mohammed & Woods, John. (2013). Survey on QoE\QoS Correlation Models For Multimedia Services. International Journal of Distributed and Parallel systems. 4. 10.5121/ijdps.2013.4305.

[91] Joe Breen, Andrew Buffmire, Jonathon Duerig, Kevin Dutt, Eric Eide, Mike Hibler, David Johnson, Sneha Kumar Kasera, Earl Lewis, Dustin Maas, Alex Orange, Neal Patwari, Daniel Reading, Robert Ricci, David Schurig, Leigh B. Stoller, Jacobus Van der Merwe, Kirk Webb, and Gary Wong. 2020. POWDER: Platform for Open Wireless Data-driven Experimental Research. In Proceedings of the 14th International Workshop on Wireless Network Testbeds, Experimental evaluation & Characterization (WiNTECH '20). Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/3411276.3412204

[92] Platforms for Advanced Wireless Research Link: https://advancedwireless.org/, Достъпено на 27.07.2025г

[124] Alamr, Abrar, and Abdelmonim Artoli. 2023. "Unsupervised Transformer-Based Anomaly Detection in ECG Signals" Algorithms 16, no. 3: 152. https://doi.org/10.3390/a16030152

List of publications related to the dissertation

[A1] Vlahov, Atanas & Ekova, Dessislava & Poulkov, Vladimir & Cooklev, Todor. (2022). Virtualized, Open and Intelligent: The Evolution of the Radio Access Network. 10.1201/9781003360889-9.

[A2] Velyova, Vesela & Vlahov, Atanas & Poulkov, Vladimir & Ivanov, Antoni. (2024). O-RAN Based User Tracking for Emergency Scenarios. 1-5. 10.1109/WPMC63271.2024.10863025.

[A3] Kougioumtzidis, Georgios & Vlahov, Atanas & Poulkov, Vladimir & Zaharis, Zaharias & Lazaridis, Pavlos. (2022). QoE-Oriented Open Radio Access Networks for Virtual Reality Applications. 491-496. 10.1109/WPMC55625.2022.10014946.

[A4] Vlahov, Atanas & Poulkov, Vladimir & Mihovska, Albena. (2021). Analysis of Open RAN Performance Indicators Related to Holographic Telepresence Communications. 1-5. 10.1109/WPMC52694.2021.9700477.

[A5] Mihovska, Albena & Vlahov, Atanas & Poulkov, Vladimir. (2024). 6G-based Intelligent, Context-Aware, and Trustworthy User-Centric Healthcare Applications. 1-6. 10.1109/WTS60164.2024.10536689.

[A6] Evgenieva, Evgeniya & Vlahov, Atanas & Ivanov, Antoni & Poulkov, Vladimir & Manolova, Agata. (2025). A Comprehensive Survey of 6G Simulators: Comparison, Integration, and Future Directions. Electronics. 14. 3313. 10.3390/electronics14163313.

[A7] A. Vlahov, V. Poulkov, P. Lazaridis and Z. Zaharis, "A Machine Learning Methodology for Network Anomalies Detection in O-RAN Networks," European Wireless 2023; 28th European Wireless Conference, Rome, Italy, 2023, pp. 174-178.

[A8] Georgieva, Polya & Vlahov, Atanas & Mfondoum, Roland & Poulkov, Vladimir & Zaharis, Zaharias. (2025). Informer-Based Anomaly Detection in Mobile Networks Using CDR Time-Series Analysis. 1-4. 10.1109/ICEST66328.2025.11098317.

[A9] Gotseva, Nikol & Vlahov, Atanas & Mfondoum, Roland & Ivanov, Antoni & Poulkov, Vladimir. (2025). A Comparative Analysis of Anomaly Detection Techniques in Cellular Data. 1-5. 10.1109/ICEST66328.2025.11098230.

[A10] Kougioumtzidis, Georgios & Vlahov, Atanas & Poulkov, Vladimir & Lazaridis, Pavlos & Zaharis, Zaharias. (2024). QoE Prediction for Gaming Video Streaming in O-RAN Using Convolutional Neural Networks. IEEE Open Journal of the Communications Society. PP. 1-1. 10.1109/OJCOMS.2024.3362275.

[A11] Kougioumtzidis, Georgios & Vlahov, Atanas & Poulkov, Vladimir & Lazaridis, Pavlos & Zaharis, Zaharias. (2023). Deep Learning-Aided QoE Prediction for Virtual Reality Applications Over Open Radio Access Networks. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3343846.

[A12] Gotseva, Nikol & Vlahov, Atanas & Poulkov, Vladimir & Manolova, Agata. (2024). ML-Driven Prediction of QoS in C-V2X Scenarios. 1-4. 10.1109/ICEST62335.2024.10639683.

[A13] Georgieva, Polya & Vlahov, Atanas & Poulkov, Vladimir & Manolova, Agata. (2024). A Machine Learning Approach for Network Slice Selection. 1-5. 10.1109/ICEST62335.2024.10639750.

[A14] Vlahov, Atanas & Kougioumtzidis, Georgios & Mihovska, Albena & Poulkov, Vladimir. (2022). Performance Analysis of Evolved RAN Architectures with Open Interfaces. Journal of Mobile Multimedia. 10.13052/jmm1550-4646.19112.

TECHNICAL UNIVERSITY OF SOFIA
FACULTY OF TELECOMMUNICATIONS
DEPARTMENT "COMMUNICATION NETWORKS"

Atanas Vlahov, M.Sc.

Machine Learning for QoE Enhancement in Future Wireless Networks

ABSTRACT of Ph.D. THESIS

Future cellular networks are expected to support the diversified requirements of emerging services, with a strong focus on applications that are highly critical to Quality of Service (QoS) and Quality of Experience (QoE), such as autonomous mobility, industrial automation, and highly immersive multimedia. This broad and heterogeneous spectrum of scenarios necessitates the development of flexible, scalable, and programmable networks capable of guaranteeing ultra-high reliability and low latency, ensuring uncompromising quality of service and quality of experience levels for end-users in a dynamic environment.

The main objective of this dissertation is to propose a network management methodology specifically targeted at QoS-critical applications. The implementation is based on leveraging the advantages of machine learning and the Open Radio Access Network (Open RAN) architecture to improve network efficiency, flexibility, and automation. More specifically, the proposed framework focuses on the development of artificial intelligence algorithms integrated into the RAN Intelligent Controller (RIC) to provide autonomous radio resource allocation, precise Network Slicing, as well as continuous QoS monitoring and forecasting. This ensures strict adherence to the network requirements of critical applications and prevents service degradation in real time.