UNIVERSITY OF THESSALY

Electrical & Computer Engineering Department

Master Thesis

"Design and Implementation of a Route Scheduling Mechanism for Connected Vehicles, subject to the performance of 5G services running on top."

Διπλωματική εργασία

Σχεδιασμός και Υλοποίηση μηχανισμών Δρομολόγησης Συνδεδεμένων Αυτοκινήτων παρουσία υπηρεσιών δικτύου 5ης Γενιάς

Author: Christos Nanis

(christosnanis@outlook.com)

Supervisor

Athanasios Korakis

Committee Members

Antonios Argyriou

Dimitrios Bargiotas

A thesis submitted in partial fulfillment of the requirements

for the degree of Master in

Science and Technology of Electrical and Computer Engineering

Volos, Greece

September 28th 2024

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελούν αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλουν οποιασδήποτε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχουν έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο Δηλών

Χρήστος Νάνης

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

The Declarant

Christos Nanis

Περίληψη

Η παρούσα μεταπτυχιακή εργασία ασχολείται με την μελέτη, σχεδιασμό και υλοποίηση μηχανισμού υποβοήθησης πλοήγησης αυτοκινούμενων χρηστών χρησιμοποιώντας πηγές πληροφορίας πραγματικού χρόνου και εργαζόμενοι στην άκρη (edge) ενός σύγχρονου δικτύου συνδεδεμένων αυτοκινήτων (V2X). Τα σύγχρονα δίκτυα τηλεπικοινωνιών αφοσιώνονται με την διαρκή εξέλιξη και επανακαθορισμό νέων δυνατοτήτων και χρησιμοποιούμενων τεχνολογιών που προάγουν την βελτιστοποίηση της διαχείρισης πόρων όπως η ενέργεια και υπολογιστικοί πόροι που απαιτούνται για την εκτέλεση κομβικών διεργασιών. Αυτό γίνεται μέσω χρήσης διαφόρων αυτοματισμών και χρήσης εξελιγμένων μαθηματικών εργαλείων όπως η Μηχανική Μάθηση για να επιτευχθεί η διασφάλιση των εγγυήσεων σε θέματα απόδοσης και χρόνου απόκρισης δικτύου πάνω στις οποίες αναπτύσσονται πληθώρα εφαρμογών όπως και αυτές που συσχετίζονται με τα δίκτυα συνδεδεμένων αυτοκινήτων. Αρχικά, παρουσιάζεται μια επισκόπηση των εξελίξεων στα κινητά δίκτυα από την πλευρά των υιοθετούμενων τεχνολογιών και γενικότερα των πρόσφατων κατευθύνσεων δραστηριοποίησης από τους τεχνολογικούς παρόχους, όπως επίσης και των κρισιμότερων προβλημάτων που προκύπτουν από την χρήση τους περιλαμβάνοντας το πώς αντιμετωπίζονται στην βιβλιογραφία. Συγκεκριμένα, διερευνάται πώς η Υπολογιστική Νέφους στην Άκρη του Δικτύου επιτρέπει τη δημιουργία κλιμακούμενων και αποδοτικών υποδομών για δίκτυα οχημάτων, όπως επίσης και αναλύεται ο ρόλος της μηχανικής μάθησης στη βελτιστοποίηση των κινητών δικτύων. Η ταυτόχρονη υποστήριξη χρηστών με διαφορετικές απαιτήσεις πάνω από ένα πλήρως δυναμικό δίκτυο ενώ υπάρχει ανοιχτό το ενδεχόμενο (υψηλής) κινητικότητας τους, οδηγεί σε περαιτέρω θέματα όπως επανασυνδέσεις σε κεραίες, αναγκάζοντας το δίκτυο να πλαισιωθεί από διάφορες αυτοματοποιήσεις και προληπτικούς μηχανισμούς αποφάσεων για την διασφάλιση της προτυποποιημένης απόδοσης του για κάθε χρήστη χωρίς να θέσει σε κίνδυνο την ακεραιότητά του. Η συγκεκριμένη δουλειά προτείνει επίσης μια σύγχρονη λύση στο θέμα πρόβλεψης των διακυμάνσεων της ποιότητας εμπειρίας ενός αυτοκινούμενου χρήστη, προβλέποντας τους δείκτες ποιότητας υπηρεσίας (QoS) στον χρόνο μέσω χρήσης τεχνικών Βαθιάς Μάθησης (Deep Learning) ώστε να εντοπιστούν προληπτικά εξασθενήσεις σήματος και να επηρεαστεί ο χρονοπρογραμματισμός των χρηστών δυναμικά από τον μηχανισμό πρόβλεψης. Τέλος, η ενσωμάτωση των προηγμένων αυτών τεχνολογιών αναλύεται ως προς το αντίκτυπο που έχει ως προς τους πόρους του δικτύου και πώς οι ανάγκες αυτές κλιμακώνονται με την αύξηση χρηστών.

Λέξεις-κλειδιά:

Δρομολόγηση Ταξιδιού, Υπολογιστική Νέφους, Άκρη του Δικτύου, Χρονική Πρόβλεψη, Βαθιά Μάθηση, Ποιότητα Εμπειρίας

Abstract

The present master thesis involves the study, design and implementation of a travel assistance mechanism for vehicles by leveraging real-time information sources at the Cloud Edge and on considering a modern Vehicle-To-Everything (V2X) scenario. It stands that the telecommunication sector is deeply involved in the continual developing and redefining of capabilities and utilized new technologies that target optimization of resource management like energy and computational resources for performing crucial network procedures. This is achieved through the use of various automations and advanced mathematical tools like Machine Learning models for meeting guarantees on matters of performance and network response time over which a multitude of applications like critical road safety build upon. First, an overview on the recent developments in mobile networks from the scope of utilized technologies and relevant research directions by the technology vendors is presented alongside critical challenges that surface from their standard use and how these challenges are addressed in peer works. Specifically, the impact of Cloud Computing at the Network Edge on forming scalable and efficient infrastructure for vehicular networks is investigated, while the role of Machine Learning systems in optimizing modern mobile networks is taken into consideration. Concurrently supporting multiple users under different requirements over a fully dynamical network environment while there is open opportunity for their (high) mobility, results in additional challenges such as handovers. This creates the need for accompanying the network with various automations and proactive decision-making mechanisms that can guarantee standardized performance without jeopardizing the safety of end users. This work also proposes a modern solution to the problem of predicting vehicle user Quality of Experience (QoE) by forecasting relevant signal quality indicators (QoS) with Deep Learning methods, so that signal quality degradation can be early detected and dynamically affect the scheduling of the users by the network. Last, the integration of these advanced technologies is analyzed in terms of the resulted impact on computational resources and how these requirements scale along with an increase in the number of end users.

Keywords:

Journey Routing, Cloud Computing, Network Edge, Time Forecasting, Deep Learning, Quality of Experience

Περιεχόμενα

List of Figures

List of Tables

Chapter 1 – Introduction on Modern Mobile Networks

1.1 Overview of Developments in Modern Mobile Networks

A new generation (G) in the mobile network development continuum usually reflects the emergence of specific market needs for new features and technologies that can meet system performance requirements for trending applications and upcoming technology vendors. Specifications for these technologies may require features like ultra time-sensitive interactions between end users and the network or massive-scale of deployment, while the battery life and computational capabilities of the end user devices form specific needs in managing and servicing user transmissions effectively.

The 2nd generation GSM architecture that was introduced in the early 1990s utilized technologies able to reach bandwidths in the scale of a few Kbit/s, while utilizing mainly analog systems that were able to provide SMS texting and digitally encrypted communications with a channel bandwidth that ranged up to 2 GHz. Next, the 3rd generation architecture introduced General Packet Radio Service (GPRS) technologies in order to complement circuit-switched technology with packet-switching. This technology standard brought upon QoS parameters for reconfigurable point to (multi-)point communications mainly for outdoor urban and rural sites possibly assisted by satellite.

Later mobile network generations like LTE focused on one business ecosystem such as mobile broadband, all while having a monolithic system design. The network architecture and specifically the base station is later disaggregated into Central Units (NAS, RRC and PDCP protocols) and Distributed Units (RLC and downward layers of the protocol stack). Functional splits are used to attain technology-agnostic, multi-tenancy support. These technologies may belong to either *3GPP* (e.g. LTE/4G), *non-3GPP* (e.g. *WiFi, Bluetooth*) technologies. These technologies require scheduling of end-to-end transmissions into achieving swiftly adapted and efficiently utilized radio resources, while transmission times are upper-bounded by restrictive values as in the case of LTE (~1ms). *Transmission Time Interval* (TTI) is defined as the length of time required to transmit a transmit block. At each TTI, a Base Station Scheduler typically proceeds with considering the Physical Radio Environment of each user, which in turn report their perceived QoS to the Scheduler that is tasked with deciding on the Modulation and Coding Scheme that is utilized for both upstream and downstream tasks.

Newer generations aim by design to effectively support multiple service endpoints for users who may also be able to be offered content for multiple applications under different needs and tightly integrated radio technologies (like 5G/6G interfaces, Wi-Fi, Bluetooth, etc.) belonging to ranging service needs, deployment vicinity sizes and frequency bands.

1.2 Main Modern Mobile Application Types

Enhanced Mobile Broadband (eMBB) applications are related to human-centric use cases that involve access to multimedia content, services and data with leveraged performance and seamless user experience. Applications belonging to this class can provide the basis for systems that come with different capabilities in terms of load, supported mobility and network traffic capacity.

Ultra-Reliable and Low-Latency Communications (URLLC) place strict requirements on end-to-end latency, reliability and availability with possible use cases including remote medical surgery, distribution automation in a smart grid, transportation safety, V2X [1].

Massive Machine-Type Communications (MMTC) are characterized by large deployments of cheaper devices with long battery life that transmit low volumes of non-delay-sensitive data, usually related to agricultural applications, smart-metering and logistics.

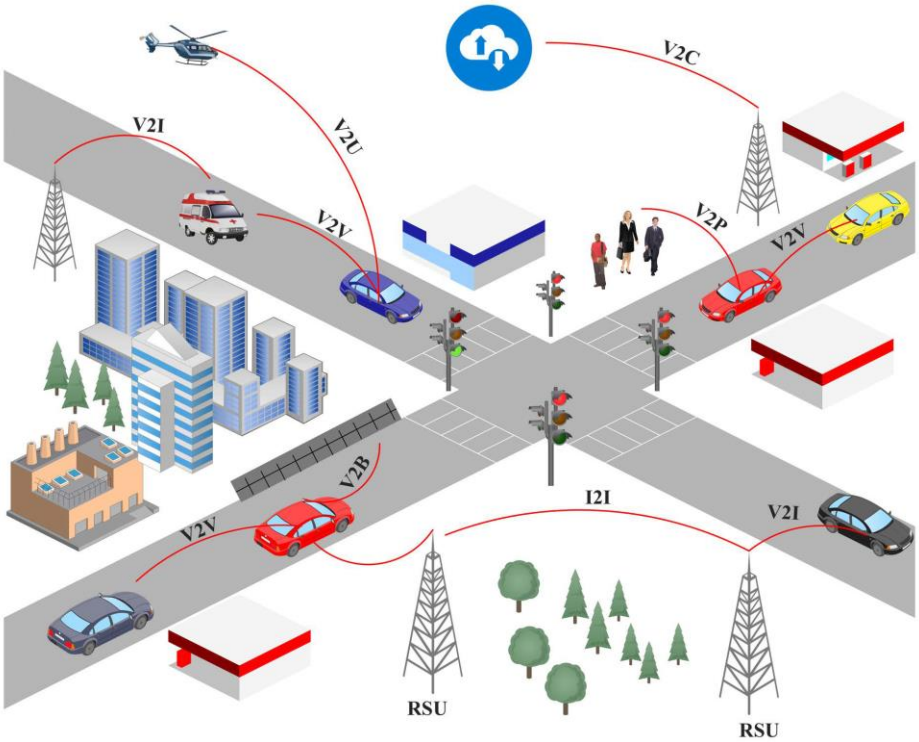

Figure 1. Overview of a V2X Network [2]

1.3 On Vehicular Networks

The Telecommunications and Wireless Networking sectors are regarded as committed to their continuous evolution by iteratively providing new generations of standards and implementations that aim to reinvent and redefine the modern network technological landscape, while providing increased support for use cases such as connected vehicles by accommodating vehicle mobility and fluctuating vehicular radio environment dynamics. Connected vehicles are increasingly equipped with various advanced on-board sensors such as GPS modules, LIDAR sensors and MIMO antennas (Figure 2). Techniques like proactive resource management target efficient support of use cases such as semi- or fully automated driving, all while generating increasing volumes of measurements. Auxiliary data measurements on road characteristics in terms of structure (e.g. road segmentation) and conditions (e.g. hazards, congestion) may be communicated from the network to the end user (vehicle-to-infrastructure/V2I link) or between vehicles (vehicle-to-vehicle/V2V links) (Figure 1).

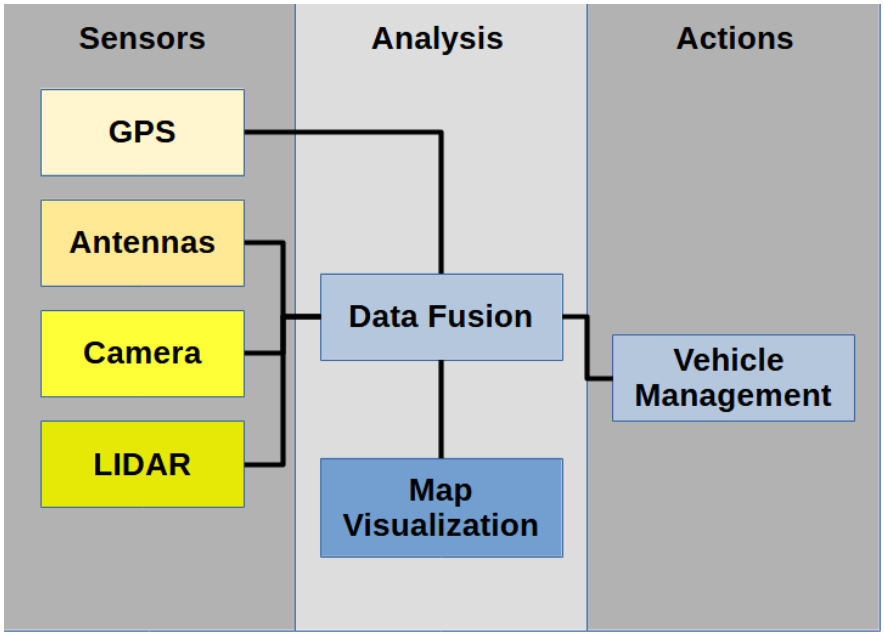

Figure 2. On-board vehicle sensor system example

In order to attain new levels of transport safety and to integrate useful Infotainment Services, Intelligent Transportation Systems (ITS) aim to provide fertile ground for applications that build on road safety and infotainment. Under such systems, vehicles are required to transmit periodic synchronization broadcast signals such as identification, location or additional hardware readings

to neighbouring vehicles and/or infrastructure at intervals that may range from a few minutes to a few milliseconds. Radio Access Technologies may generally be heterogeneous (e.g. 802.11 channels, LTE/4G, 5G/B5G). Connected vehicles may leverage established connections to services, on-board hardware and sliced cloud resources into performing assigned tasks. Further optimizations to this concept may include the formation of autonomous fog networks between vehicles for sharing free-floating resources. Key concepts in the successful integration of cloud-native technologies to vehicular networks include:

- Network System Resilience: Recovering from failures that owe to vital component malfunction and general connectivity or service access disruptions is considered of paramount importance to post-4G network deployments that promise all-connectivity irrelevant to location and mobility.
- Ultra-Low End-to-End Latency: The incorporation of cloud and edge computing techniques facilitates localized data processing within single-hop distances, resulting in considerable reduction in round-trip times that can meet restrictive requirements for e.g. road safety applications.
- Modularity of Software: In order to provide a fertile ground for vehicle management and road safety applications that may build on top, operations related to on-demand scaling and dynamic service adaptation should be able to be applied without degrading overall system performance.
- Data Analytics and Artificial Intelligence: Advanced data processing techniques and artificial intelligence (AI) may be leveraged by the network in uncovering data-driven semantics and associations to aid orchestrating automated decision-making network operations for vehicle management, road/traffic congestion mitigation and travel itinerary generation.
- CI/CD Best Practices: Continuous integration/Continuous delivery methods can back the agile development, prototyping and deployment of in-network software and services when working with cloud-native computing infrastructure. Networks can frequently update their capabilities and perform crucial updates while remaining adaptable and operationally efficient.

1.4 Mobile Network Environment and User Mobility (Handovers)

As mobile data keeps being accumulated to complement modern big-data workflows, new methodologies are brought in to satisfy specified performance guarantees (e.g. QoS requirements) for crucial in-network services. Mobility patterns by connected user devices can fundamentally

impact radio network environment dynamics by naturally roaming in-between, out and around the vicinities of base stations under non-uniform radio access technologies and transmission configurations. Despite this fact, highly mobile users in the network need to be able to efficiently perform context information transfer including accumulated radio/application session states to another cell which is capable of continuing to provide service for the user after the quality of experience for the latter starts to degrade significantly.

Traditional control algorithms are mainly driven by heuristic methods that are largely dependent on expert-tuned solutions deemed to be slow and expensive. These procedures are UE-assisted and network-controlled [3] in that the user equipment is responsible for distributing QoS measurements to the network infrastructure, which in turn can perform the handover decisions. First, handovers can be classified by use case such as *quality-based* handovers that are triggered by UE-reported QoS measurements that prove non-SLA satisfying. Such procedures may be triggered by requesting a seamless switch to a suitable cell for smoothly continuing current user network session. *Coverage-based* handovers form another class, triggered by specific user mobility patterns that may include moving away from a given cell vicinity (inter-RAT type of switch).

According to another taxonomy that does not require a reference signal by the users, an *interference-aware* handover decision can enable switching to a smaller cell in a multi-technology network, within which multi-layer interference can account for cell site conditions together with Received Signal Quality (RSQ) measurements at the user side. *Speed-based* handover decision algorithms may compare a given user's speed against a specific threshold to mitigate problems with handover execution, especially for high-speed users that require rapid handover execution. Further, by incorporating relevant information on network load and current traffic levels, relevant methods can manage decrease in overall handover signaling costs. Lastly, a *load-based* handover decision algorithm would consider the service access latency that comes with the user experience.

Chapter 2 - Cloud Computing as Enabler to Mobile Networks

Cloud-based mobile applications have skyrocketed in demand through recent years. Applications such as real-time image recognition and video streaming impose restricting requirements for the end-user equipment i.e. requiring high data processing capabilities. While market-distributed smartphones experience continuous improvements in specified capabilities, their limitations concerning battery life and processing power render them as a last resort for executing tasks of high complexity. Cloud Computing techniques have thus been proven to be useful for users, while being responsible for carrying out what is needed by network users of a demanding application.

2.1 Cloud-Native, Scalable Software Development

Cloud-native computing is a modern approach to software development which leverages cloud infrastructure capabilities in order to design, deploy and manage scalable applications within dynamic environments and variable runtimes. Building and running applications that instrument distributed computing capabilities that are offered by the cloud delivery model can greatly impact the cloud capabilities in scaling, failover, elasticity, etc. This modus operandi utilizes advanced technologies such as container processes managed by serverless or server-based architectures and services to enable different types of modern cloud deployments in public/private or hybrid cloud environments. Virtualization techniques may specifically be used towards full automation and standardization of algorithms and network functions within diverse environments.

Cloud-native architectures are known to primarily tackle operational complexity in order to efficient manage applications that can scale with minimal computational overhead. This design may further grant abilities in fine-grained monitoring, analyzing end-to-end performance, advanced capabilities in diagnosing and troubleshooting problems that pertain to per-component or system performance. Applications that are developed under this paradigm can be found to predominately follow an architecture that involves one or multiple application containers. Managing these containers at scale can further employ orchestration platforms such as Kubernetes in order to automate deployment and auto-scaling while enabling efficient resource management. Alternative open-source implementations can also be found to support cloud-native development.

By bringing cloud-native principles into the system, organizations may reach significant improvements in operational efficiency and agility, a vital feature in meeting evolving cloud environment demands for various use cases and business verticals.

2.2 Edge Computing and Processing Units Placement

Cloud structure typically follows a centralized structure, imposing a large geographical gap between the users and provided services, contributing to an unwantedly high end-to-end communication latency (multiple in-between hops). This observation creates the need for installing infrastructure (high-processing-powered units) at the edge of the Network in proximity to the end users. Edge computing thus optimizes quality of experience by performing data processing closer to the user, thereby reducing end-to-end latency and meeting real-time deadlines. This methodology mitigates problems that are met when relying solely on a centralized cloud infrastructure for offloading computational tasks under stringent time requirements. Installed processing units are called Multiple-Access Edge Computing units, merely mobile edge clouds that render end-to-end performance as fast and manage to decrease traffic congestion that arises in the backhaul links. A cloud infrastructure that is complemented by roadside edge nodes (RSUs – Road Side Units) can thus play a crucial role in assisting communication among users and the network infrastructure by:

- Providing access to service content and resources under high availability and zero downtime through selection of access technologies.
- Deciding on future service failures and mitigating them in real-time.
- Dynamically deciding on transmission schemes that enhance user device battery life and promote energy efficiency.

As a perfect key enabler to various real-time and context-awareness-enhancing technologies, Edge Computing is expected to leverage interconnectivity and intercommunication in a highly-collaborative and highly-responsive intelligent cloud network. Mobile subscribers, enterprises and vertical segments that participate in an Edge-enabled network are provided more freedom to authorize third-party vendors with implementing custom service ecosystems. Applications that leverage edge-enabled computation may have to do with use cases that loosely relate to computational offload, local connectivity and caching, content optimization [4].

2.3 Cloud-Enabled, Edge Computing-assisted Vehicular Networks

Coupling Cloud infrastructures with edge computing capabilities is expected to play a catalyst role in modern mobile networks especially for emergent vehicular networks. This allows computational tasks to be efficiently distributed between centralized cloud infrastructure and localized edge nodes. A cloud infrastructure may specifically be tasked with providing needed basis for auto-scaling resources for large-scale computations, while the edge computing units at the side of the road may swiftly handle latency-sensitive tasks by performing tasks like data processing closer to the vehicle. Modern vehicular network applications form increasingly demanding profiles in real-time decision-making capabilities and various network operations that need to be dealt with efficiently without sacrificing performance.

2.3.1 Cloud-Based, Edge-assisted Applications in Vehicular Networks

Connected vehicles to a cloud network infrastructure may accumulate data measurements into massive quantities subsequently fused with measurements that belong to other vehicles, as needed by various time-sensitive processing tasks and analytics. Cloud/Edge computing is thus understood to be integral in this endeavor by enabling *real-time data processing* and *analytics* via acquisition and analysis of vehicle measurements at cloud servers and roadside infrastructure. This information is utilized to provision for various control operations and to produce accurate inferences for matters such as network capacity, network traffic, identifying road blockage, detecting environmental changes, etc. Transportation system infrastructure systematically offloads such tasks to the cloud and edge processing nodes to optimize and fine-tune system performance and maintain network operational efficiency.

Besides time-sensitive analytics, a cloud infrastructure may form histories of past measurements that may be later submitted to processes that perform data exploration, analysis and pattern recognition in order to orchestrate crucial in-network operations. Insights from these procedures may be leveraged by various routines that target predictive management and operational cost optimization. By further supporting advanced vehicle features such as access to infotainment and multimedia services, cloud computing may also leverage vehicle capabilities in multimedia and entertainment for passengers as in high-definition video streaming applications which specifically require ample user bandwidth and a moderate use of computational resources.

Figure 3 showcases a fully-fledged, multi-technology network architecture variant complemented by Edge Computing units that can manage and provide support for various service types:

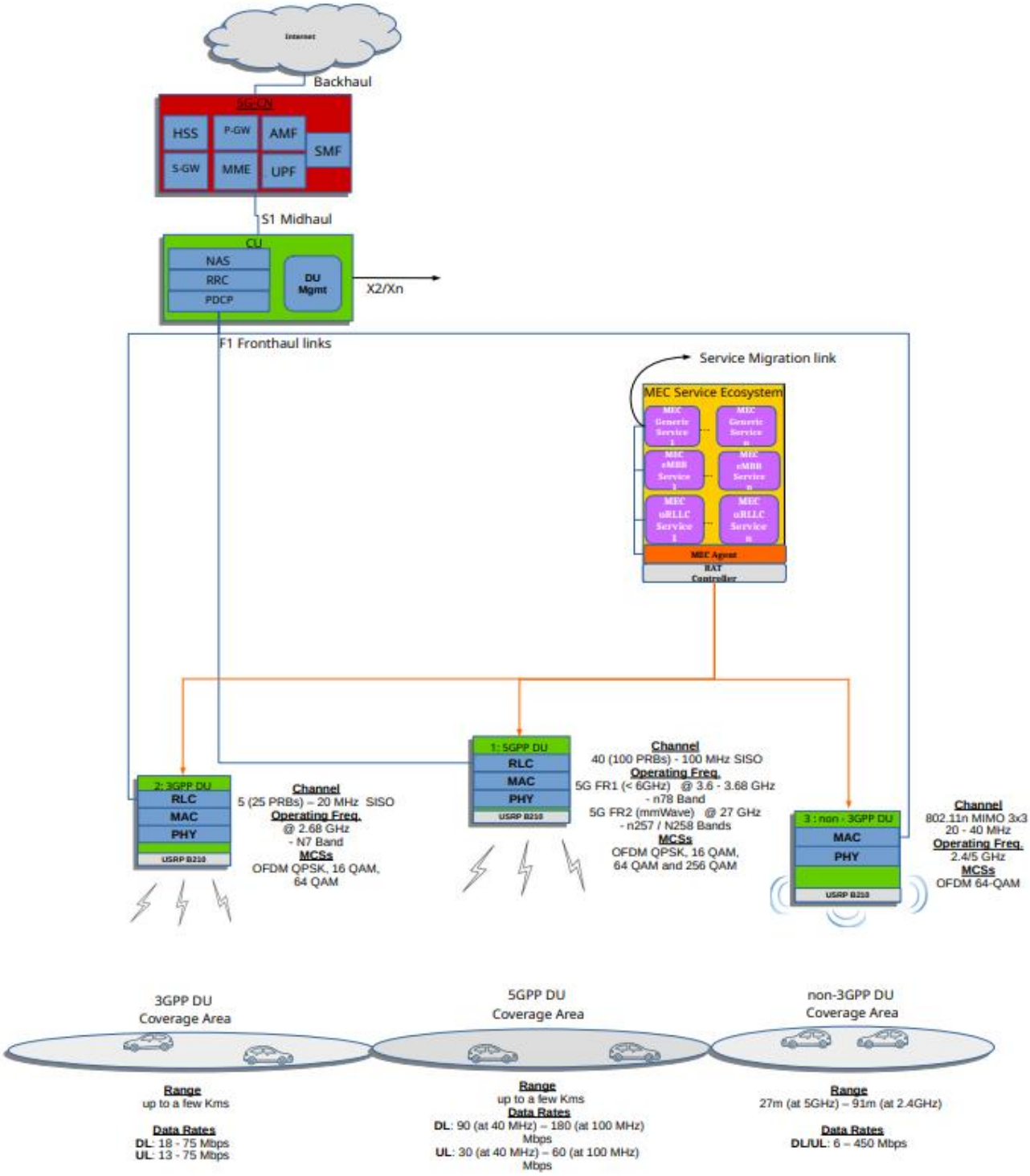

Figure 3. Overview of a multi-technology V2X Cloud Network with Edge Computing Units

2.3.2 Benefits of Cloud-Enabled Vehicular Networks

Advanced features such as dynamically allocating/adjusting cloud resources and on-demand scaling can be considered as critical advantages of employing a Cloud Computing architecture, which is able to limit performance degradations in e.g. service quality while providing efficient scaling capabilities that boosts research areas such as smart analytics and network congestion mitigation parallel to the constant growth in data measurements and number of connected vehicles. For road safety, the cloud infrastructure plays a key role in instrumenting the coordination of vehicle movements, improving system load balance and in efficient transportation system management. It follows that supporting advanced applications like road safety for network-assisted vehicles means that real-time decision-making and data processing are imperative to user safety and system component performance. Cost efficiency can further be considered an advantage of deploying cloud environments. As mentioned before, this can be achieved by offloading computations to cloud infrastructure when faced with demanding tasks, leading to cutting back on end device energy consumption while loosening the need for expensive high-process-powered vehicle on-board hardware.

Chapter 3 – Overcoming Challenges in Cloud-Native Support for Vehicles

Recent developments in the telecommunications sector have led to design paradigms that may separate Control and User planes (CUPS), bring high-process-powered roadside nodes near the user (MEC), orchestrate virtualized network functions (NFV) or define multiple layers of networks purely by software (SDNs). Moreover, matters such as slicing network resources for users during challenging wireless network conditions and intricate network traffic patterns need to be addressed. Allocated resources may refer to a multitude of established connections, storage capabilities and computation primitives all while aiming for greater spectrum efficiency, failover ability and on-demand scaling abilities. In adopting aforementioned technologies, especially NFV, network users can be offered finely-sliced resources and access to virtualized network operations under a modular service design and advanced capabilities in dynamic service adaptation according to shifting network conditions and end user device heterogeneity.

3.1 Radio Environment Non-Determinism

Vehicle mobility-induced patterns render collecting location information as paramount in facilitating the communication and coordination of involved devices amongst end users and the infrastructure. Frequent location measurements are thus deemed as essential to a variety of applications, including critical time-sensitive use cases such as road safety applications. The ability for a vehicle to receive network/vehicle-initiated messages on road hazard and traffic information in order to perform well-timed decisions that are informed for and may also consider current and/or future network state as predicted by advanced mechanisms are considered key to future in-network intelligence.

A vehicle may follow non-deterministic trajectories under varying speeds, causing rapid fluctuations in the signal quality. High mobility is known to disrupt signal quality measurements, often at the point of attachment to a new cell. Furthermore, environmental factors such as presence of obstacles, tall buildings or natural terrain and effects can profoundly damage user experience. Difficulties in predicting such patterns can create connectivity issues that result in unreliable transmissions and exacerbated latency. Further, implementation of real-time applications which build on all-connectivity as in the case of navigation and safety systems is hindered from full

fruition, if error-handling mechanisms and the edge computing principle are not adopted. On the other hand, assuming that mentioned methodologies are followed, the network system can exhibit increased resilience to failures, enhanced performance and decreased difficulty in meeting specifications for system performance.

Methods like reconfigurable radios and dynamic spectrum management under software-defined networking (SDN) middleware are expected to provide further assistance in meeting performance requirements. SDN decouples control from the hardware components, allowing for granular control over network resource management and easier real-time adaptation to variable network conditions. Dynamic spectrum management involves readjusting settings (e.g. utilized frequency bands) and deciding on optimal interference levels for a subsequent transmission. In this context, cognitive radios can automatically detect optimal subsets of available spectrum, within which adjusted transmission schemes may mitigate high interference levels. Outlined techniques can thus efficiently adapt to environmental changes, reducing the impact of non-deterministic behaviour. By further employing real-time monitoring of network conditions, mentioned systems can attain to ultra-reliable transmissions, optimize end-to-end system performance e.g. in latency and finally maintain quality of experience that does not violate service-level agreements.

By stimulating the collaboration of devices on the edge to offload computational and communicational tasks, the link connectivity and Quality of Service indications in Cloud/Edge-enabled vehicular networks can be leveraged. Vehicles may serve as edge nodes, contributing to real-time management of traffic by helping in minimizing the average response time of the reported events such as traffic jams, road hazards/accidents, etc. Vehicles may record images, videos or generally informative messages that can be uploaded using the technologies available and depending on the policy imposed for technology selection. V2X communications can be further enriched with value-added services, such as car parking space tracking and Infotainment (e.g. Video Distribution). Based on the GS MEC 012 – specified Radio Network Information Service (RNIS) which shall provide actual RAN information related to end user equipment [5], extensions aim at providing V2X applications with predictions on the QoS performance of the V2I (Vehicle to Infrastructure) link [6].

3.2 Vehicle Management

Providing support for vehicle management under an increasing number of connected vehicles within a cloud-enabled vehicular network should typically mean that a diverse range of vehicle types each shipped with specific sets of capabilities and communication requirements for its supported features must be able to transparently function within the network. Additionally, various user preferences may add another level of abstraction on top of these specifications, posing challenges in network-wide support due to differences in utilized hardware components, firmware and software versioning, supported communication protocols, etc. Assuming a uniform support for all connected vehicles despite differences in hardware components and manufacturer specifications, dynamic resource allocation can efficiently manage fluctuating vehicle network demands by real-time adjusting computing, storage, and networking resources based on live and/or historic network traffic reports.

3.3 Data Integration

System and service capabilities diversity within a cloud-native mobile network can prove challenging in efficient management, possibly requiring the presence of a unified platform that can process multi-source measurements and reduce task complexity for processing nodes in the Cloud/Edge ecosystem. By alternatively using transparent web-based APIs, vehicles and infrastructure may transparently pass messages under a technology-agnostic style that permits system reconfiguration abilities without resorting in coordination with specific vendors or systems in case of troubleshooting.

3.4 Security and Privacy Implications

Protecting sensitive information like vehicle location data and other hardware readings is crucial for cloud-enabled vehicular networks. Vulnerabilities such as the possibility of a data breach or eavesdropping points to employing encryption (i.e. data scrambling against unauthorized access) and obfuscation (i.e. creating additional levels of interpretation) techniques to mitigate these critical security failures. Access control can be considered as another important enabler for secure data handling. This technique defines a group of legitimate users and services that are allowed to view, handle or edit sensitive in-network content, thereby mitigating information abuse.

Frequently auditing access patterns in the system can lead to a timely detection of intruding entities that can prevent information loss and compromised system components. Other available methods that can increase the level of data security is local handling (e.g. in-vehicle) and processing of sensitive measurements, avoiding transmissions to other sites in the infrastructure where vulnerabilities may arise. Further, stripping off uniquely identifying fields of transmitted messages (i.e. data anonymization) can further leverage in-network data privacy. Aforementioned privacy-preserving techniques are expected to leverage in-premise multi-party security and confidentiality.

3.5 Network Intelligence

Cloud-enabled networks can contain vehicles bearing a multitude of sensors that can generate vast amounts of data points according to their followed trajectories and accessed services. Network intelligence thus refers to leveraging data-driven insights in order to enhance end-to-end performance and further assist decision-making in the network system. Advanced analytics and machine learning algorithms are possible directions to draw from in order to identify patterns and predict network/user behaviour towards intelligently reconfiguring end-to-end transmissions, optimizing resource allocation, orchestrating service access. Maintaining a seamless integration with central cloud services while offloading tasks such as data processing and advanced analytics to the edge cloud can result to full fruition from the combination of network intelligence and cloud/edge-enabled network systems.

Chapter 4 - Machine Learning-Enabled, Cloud-Native Mobile Network Applications

Recent advances in Machine Learning (ML) and Artificial Intelligence (AI) promise to leverage vast amount of data points accumulated over time at end devices and modern cloud servers in order to provide the ability of forming advanced decision-making processes under a data-driven workflow. These intricate algorithms are formed by iterative, learning-based procedures over state-of-the-art mathematical models to create systems capable of performing complex tasks such as natural language processing (NLP), image recognition, time series forecasting, etc. ML is mainly focused in uncovering associations between elements of the feature space. These associations can either be filtered or upscaled utilizing an architecture that can iteratively yield a prediction output based off currently observed data. The telecommunications sector has already been associated with specific state-of-the-art technologies that may involve a combination of cloud-native computing technologies and ML systems mainly for use cases such as urban and rural smart grids and semi-autonomous driving safety applications.

The fusion of mentioned technologies attempts to optimize network performance and automate key in-network operations. For example, tasks such as forecasting future network load or managing network traffic can be automated, thereby improving cloud resource management significantly. The ability to dynamically reconfigure resource allocation as in user/system bandwidth can save energy (e.g. in wattage), which is considered an increasingly relevant and important consideration to take into when maintaining sustainable balance between technological prowess and efficiency without jeopardizing the environment.

Most importantly, ML-aware settings can improve the quality of the user experience in terms of e.g. end-to-end latency or bandwidth by automating maintenance operations such as service failure detection and recovery. Use cases such as Narrowband Internet of Things (NB-IoT) can greatly benefit from the use of ML to achieve seamless connectivity for a wide range of power-constrained devices and be operationally efficient under ample connectivity. This work investigates the combination of technologies that can provide needed transition to advancements including the Internet of Things (IoT) and time-critical 5G/B5G into servicing increasing demands from infrastructure as in smart city or autonomous system deployments. Some of the major

applications that consider a modern cloud/edge infrastructure together with ML/AI-enabled components to profoundly impact the telecommunications industry will be showcased next.

4.1 Data-driven Network Functions

The vast accumulation of data points from various sensor hardware located in on-board vehicle sensors, roadside infrastructure or centralized cloud servers surfaces the need for intelligent data analytics and processing by AI-powered mechanisms that can predict future network conditions, optimize spectral efficiency and leverage road safety for highly mobile users. In the latter case, advanced algorithms can organize informing connected neighbouring vehicles of road blockage or malfunctioning vehicles and predicting the possibility of a road hazard to occurs in the next time window. Vast amounts of data necessitate a highly efficient data acquisition mechanism that does not jeopardize system stability by overcommitting resources. Various hardware readings that may include on-board sensors, GPS, LIDAR and camera feeds may enter the cloud in large quantities, while sensible management of which consists of tight integration with AI-enabled system control loops. AI models in cloud-native vehicular networks may allow for real-time service adaptation or transmission reconfiguration to sensibly anticipate shifting network and network traffic conditions. These mechanisms may extract useful dynamics from recorded behaviour patterns in terms of vehicular kinetic profiles and quality report measurements to perform intelligently informed decisions in resource management and route planning.

Ensuring low-latency decision-making is a major challenge in implementing data-driven network functions for these networks. Kinetic profiles are usually characterized by high mobility and radio environments by fluctuating dynamics, requiring data processing cloud capabilities that are swift enough to react to these changes. Filtering data from a potentially huge and diverse pool of sources subject to hardware errors is considered an integral component of these systems. Therefore, it is essential to apply techniques for sanitizing datasets before submitted as input to AI-driven workflows and pipelines in order to ensure the integrity and accuracy of the information at hand.

In the area of capacity forecasting for the PCF (Policy Control Function) module of the 5G/B5G Core Network, the authors of [7] design and implement a time forecasting system that can explore mobile network traffic space and time correlations using a CNN (Convolutional Neural Network) model. The implemented system utilizes centralized collection of network data for

subsequent publishing to other network management and orchestration modules. The future aggregate load is then forecast across different infrastructure locations, intelligently assisting VNF placement decisions. If forecast load reaches certain levels, then VNF demands an upscaling operation by dynamically allocating more resources. The frequency of these actions is reported to range from a few minutes to a few hours depending on the specific algorithm variation. RAN resource reconfiguration frequency may be limited by the NFV technology, so the operator must decide in advance the resources that will stay assigned to a slice until the next reallocation takes place. The authors' solution is compared in cost to traditional ARIMA-based methods without using a hyperparameter section or performing statistical forecast error measurement.

4.2 Network Traffic Optimization in Vehicular Networks

Network traffic optimization is used in vehicular networks to maintain efficient vehicle-to-everything (V2X) communications. AI/ML models deployed in cloud-native environments can monitor and predict vehicular traffic patterns, allowing for effective and spectrally efficient data transmissions. For example, if a sudden increase in vehicle density is detected in a particular area, then the system can dynamically reroute in-between hops in order to prevent network traffic congestion and ensure low-latency communication. However, one of the primary challenges in such systems is the potential delay in data transmission between the edge (vehicles) and the cloud, which may hinder real-time traffic optimization from full-fruition, especially for highly mobile users. AI-based resource management in B5G networks can assist in swiftly adaptable transmission configurations, improving the resilience and efficiency of resource management, resulting in timely decisions and optimal network performance [8]. Except optimizing data routing, cloud-native AI-powered vehicular network can perform frequent monitoring jobs that trigger resource reconfiguration to avoid performance bottlenecks and to maintain a high quality of experience in the face of rapidly changing network conditions.

4.3 Predictive Maintenance and Fault Management

The use of AI components in cloud-native networks is important in ensuring the smooth operation of both on-board vehicle hardware components and system infrastructure such as roadside units (RSUs) and cloud servers. Specifically, detecting problems such as component

failures and disconnections may allow for recovery and maintenance tasks that take place in a timely manner and do not jeopardize overall system performance. Employing a proactive approach helps avoid critical connection disruptions and leverages system reliability. Considering the high degree of vehicle dispersion and rapidly changing trajectories, it is critical that these systems detect and resolve issues swiftly. Combining measurements and results drawn from different vehicles and roadside units is essential to fault detection and mitigation, since the predictive mechanism can anticipate issues and timely address them. Examples such as hard-QoS for self-driving cars, where user awareness on the quality of experience under each accessed heterogeneous service is required, the authors of [9] suggesting forecasting future QoS values to determine whether a service is about to fail. Such functionality can be contained in a middleware to choose the optimal service at design time and conduct a dynamic service adaptation if the QoS begins to degrade, thus avoiding SLA violations. The authors report that traditional time series models require frequent retraining, while their work does not provide a hyperparameter section for utilized LSTM and GRU models.

Another study by the authors of [10] predicts uplink throughput to follow 3GPP Release 16, which allows a cellular 5G communication system to notify a V2X application of an expected (or estimated) change in QoS before it occurs, such as safely stopping a service or adapting the mode of operation for an application. Avoiding sudden session interruptions due to QoS degradation is especially important for critical V2X services (e.g., safety, automated driving). QoS sustainability analytics can therefore be provided for a specific geographic area and time interval. Training is done offline, while auxiliary features including the number of vehicles is the result of ARIMA-based simulations at inference time for the specified forecast horizon.

4.4 Energy Efficiency Optimization

The use of machine learning (ML) mechanisms is expected to significantly improve energy efficiency in communication between vehicles, roadside units (RSUs) and cloud server machines. The network system needs to optimize energy consumption for the roadside units and other system components by using algorithms that analyze network load and predict future demand, especially during periods of low network activity. This capability is especially important for smart grid systems due to power restraints and long plan of use under minimal energy consumption. In dynamic vehicular environments, it is challenging to achieve energy efficiency while maintaining reliable connectivity and high quality of experience, an intricate tradeoff that has not yet resolved.

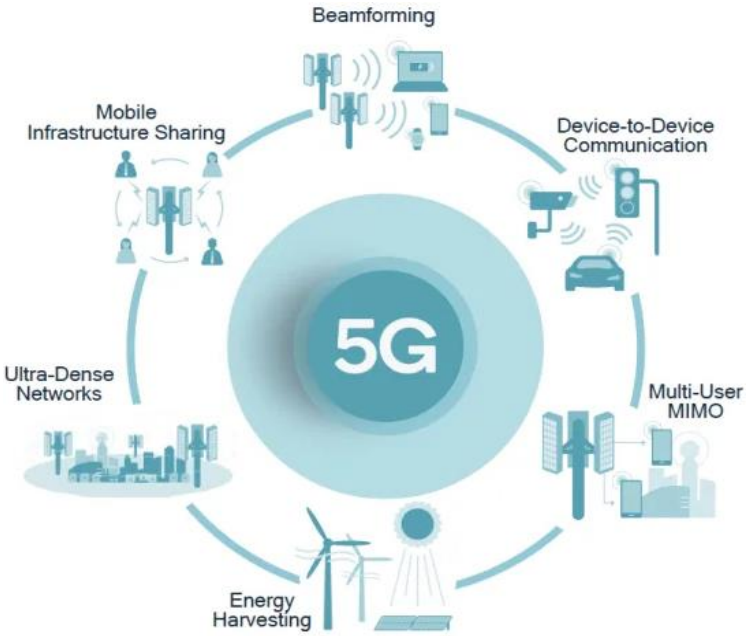

Figure 4. Energy Efficiency in 5G/B5G systems

4.5 Network Security and Threat Detection

In the wake of big-data and rapid communications, network security comes into focus especially for V2X systems, specifically their sensitive vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) links. Cloud-native, ML-enabled system components can be helpful in

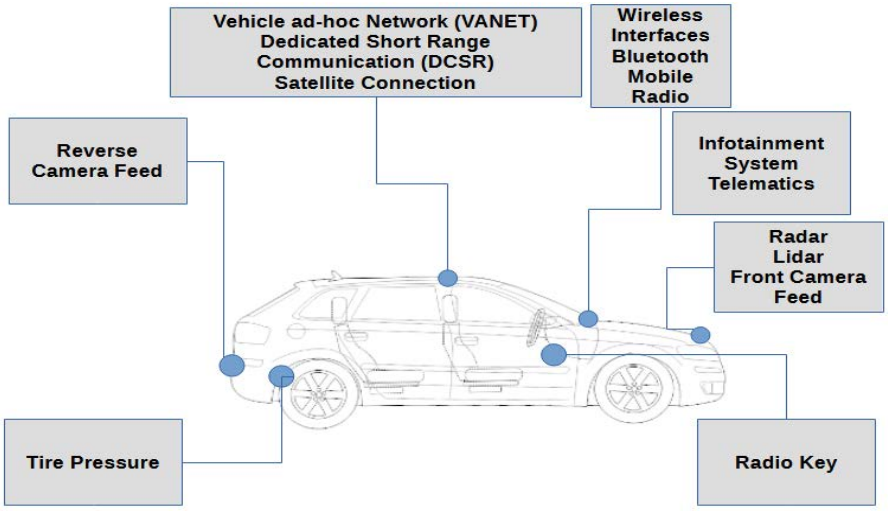

Figure 5. Vehicle Areas that contain security vulnerabilities

analyzing vast amounts of accumulated network traffic measurements in order to detect future compromise of security as in data breaches or unauthorized access, thereby ensuring communications security. Other common vulnerabilities include spoofing credentials, tampering with data and denial-of-service (DoS) attacks on vehicles or roadside infrastructure. There exist considerable challenges in real-time detecting and mitigating security threats, especially in fast-moving environments like vehicular networks. Effectively detecting and neutralizing security vulnerabilities requires delay-sensitive data processing abilities by the cloud/edge infrastructure, so as large datasets may be quickly analyzed offline, without sacrificing the ability to perform real-time decision-making based on current data.

4.6 Quality of Experience (QoE) Optimization

The Quality of Experience (QoE) for users can be optimized through the use of intelligent systems that dynamically adjust network parameters in real time, based on data collected both from vehicles and network infrastructure. ML models can prioritize critical communications to ensure that vital system resources like bandwidth and service access latency can meet design-time specifications. Maintaining consistent QoE under highly dynamic conditions such as changing vehicle speeds, varying network performance and fluctuating road conditions continues to be a significant challenge for both edge and cloud devices. A timely acquisition of measurements in terms of location, traffic congestion and fuel consumption can provide needed basis for estimating user experience with minimal deviances. The estimator module can generate predictions on future user experience based off accumulated data from various onboard and roadside sensors, real-time traffic data and historic behaviours, if applicable. This procedure may utilize limited resources that are allocated for the sole purpose of online training.

4.7 Learning-Enabled Channel Estimation

Accurate estimation of radio network channel properties in real-time consists of an advanced feature in modern wireless communication systems. As wireless networks become more intricate and densified, implications in hardware complexity need to be taken into consideration (e.g. MIMO antenna design complexity), which may hinder the ability to efficiently allocate vital resources at the transmitter to combat interference and to optimize end-to-end performance.

Vehicular radio environments can also be characterized by highly volatile Doppler shifts, short channel coherence periods, which can significantly impact transmissions and damage system reliability and end-to-end performance. For example, the authors of [11] present their design of an online training module for downlink scheduling in the 5G New Radio (NR) system to mitigate the negative impact of outdated channel quality information on communication, particularly in high-speed mobility scenarios. It is reported that for CQI prediction, a well-trained module is difficult to apply, demanding training of a new model when prediction error begins to peak. The findings of the authors emphasize the importance of speed with numerous plots depicting its impact on e.g. achievable throughput (in Mbps).

4.8 Vehicle Trajectory Prediction

Machine learning (ML) systems can utilize vehicle trajectories derived from geolocation data, vehicular kinetic profiles and environmental effects as input, in order to provide crucial predictions for autonomous driving and collision avoidance systems. These models can predict the time instance when vehicles change lanes or perform specific maneuvers (e.g. sharp turns), allowing nearby vehicles and infrastructure to appropriately respond thereby increasing road safety. However, the authors of [12] point out that achieving high accuracy in predicting vehicle trajectories is especially difficult in urban vehicular environments that contain multiple latent factors such as road hazards, traffic signals, and pedestrian movement that cannot be explicitly accounted for, making accurate trajectory prediction difficult to attain. Additionally, network latency is subject to noticeable increases, especially when processing large amounts of trajectory data in-cloud, hindering real-time applications such as road safety systems to fully utilize an ML prediction mechanism. Existing collision detection and warning systems typically use methods that either estimate the minimum distance to collision or perform assesment of risk. The former method estimates the time to accident based on vehicle kinematics and signal strength, while the latter method employs methods like probability modeling and analysis to quantify the likelihood of specific network actions that can prevent or mitigate collisions.

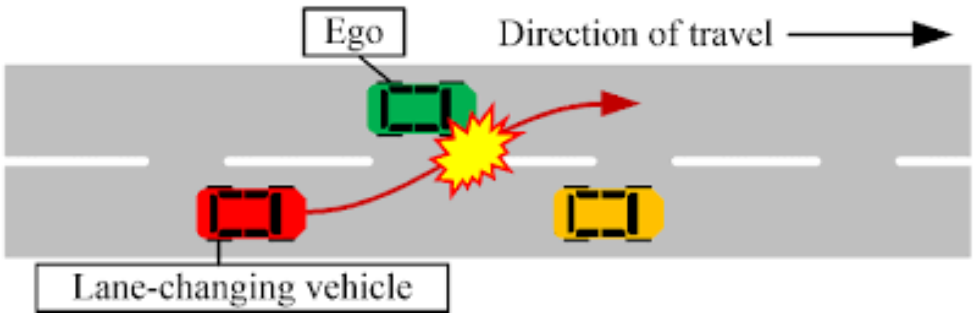

Figure 6. Illustration of a Basic Example for collision detection and avoidance [13]

Chapter 5 - Mode Of Work

5.1 Considered Model Definitions

Supervised Learning (SL) – based solutions are employed to problems of multi-step time forecasting of interrelated metrics (e.g. Channel Quality Indicators in the case of RSSI) that may belong to 3GPP (LTE) and non-3GPP Radio Access Technologies (RATs).

Early attempts involved the use of linear models (AR/ARIMA). Linear Models are estimated with unbiased (deterministic) estimators that solve for one/multiple target variables via Ordinary Least Squares (OLS). Different hyperparameter configurations for these models give rise to different properties. It is beneficial to investigate multiple configurations since e.g. two models of the same candidate model set, with one having far less parameters than the other, may possess similar test errors, us giving preference to the former in being selected as the final model used for the actual forecasts. A hyperparameter configuration may include the lag length, moving average length, transformations (e.g. first differencing), etc. With the strategy described, a set of models of different configurations are fitted to the training data and evaluated on a test set. This leverages Model Selection techniques that are associated with an indicative cost in model parameters. Experimentation showed that frequent retraining is needed for multi-step forecasting with these models, which is in agreement with various peer works that were analyzed previously.

Consequently, the sole adoption of non-linear recurrent feed-forward models (LSTM/GRU DNNs) was deemed to be necessary. Non-linear models are estimated with reduction–based iterative methods like Stochastic Gradient Descent (SGD) and variants e.g. mini-batch SGD. These models are described by a single, but far more extensive hyperparameter configuration that is proven to perform over a given number of different random initializations, enabling model forecast error analysis under statistical uncertainty. These errors are derived with the MSE loss function off a validation set of seen examples and evaluated on an unseen test set to provide a measure of out-of-sample forecast error performance.

Neural Network-based models (NNs) alter the information content at input with well-defined mathematical functions (activation functions) and operations: y = $g(W * x + b)$, where y is the model output, x the input, g the activation function (e.g. sigmoid or tanh), W the model weights and b the bias. This operation describes the work of a single unit (neuron) in the model, during the forward pass. Units scale vertically to form a layer, while scaling horizontally forms a

multi-layer architecture (a DNN), where each unit in a layer has a connection with each unit in the next layer. Model training is an iterative process that is segmented into training rounds (epochs). The course of an epoch contains a forward and a backward pass. The former was described before and the latter involves evaluating the error in the model's prediction that corresponds to time index t+h by comparing against the ground truth (actual value) which initially belongs in the validation set. The model weight updates may thus depend on the training examples, when the validation examples serve as indicatory of the model's error generalization but are iteratively used and so are prone to biased solutions. Next, the utilization of a 3-way dataset split of contiguous sets of training (seen), validation (seen) and test (out-of-sample) examples can alleviate these problems but creates the need of having collected the future ground truths beforehand to perform a simulation or coordinating their real-time collection so as the loss can be evaluated after each forecast point.

Indulgence to the opposite direction of the error surface (a quadratic bowl for an MSE loss function) is controlled by the learning rate (η) that is one of the most important hyperparameters to tune. In fact, SGD requires only a sufficiently small learning rate to operate and thus reach a possible solution. More rounds of training are not guaranteed to reach better results, since training losses are minimized to the point of a noise ball, this effect having been coined the term of *vanishing gradient*. It follows that providing a means to stop training before a maximum number of epochs is desirable and this is usually covered by *early stopping* practices. This procedure monitors the consistency of the validation errors, stopping training when the validation errors stagnate for a fixed number of consecutive training rounds. Gradient descent variants generally impose a framework for the weight updates as: $w^{k+1} \leftarrow w^k - \eta \nabla J(x, w^k)$, where w^k are the model weights at iteration k, x the input, η is the learning rate and ∇J the gradient of the loss function J (e.g. L2 error/MSE). If the gradient is based on a single training example, this is referred to as *online* Stochastic Gradient Descent, else if it is based on a batch of multiple training examples, that is referred to as *mini-batch* SGD.

5.2 Considered System Definition

This work is mainly interested in stateful network services that reside at the edge of the road and pertain to vehicular applications such as road safety and infotainment. These services are offered by high computing power nodes that reside at the edge of the network to assist in task offloading and are designed to be able to be migrated under strictly zero downtime when it is

deemed necessary (e.g. by following the trajectory of a vehicle). User state refers to established connections, control and user information, etc. The migration procedure is known to take up to several minutes, creating the need for an early detection mechanism that can attain true zero downtime of media accessed.

User quality measurements are able to reflect the view of the user under all-connectivity that is irrelevant to position, something that is supported by network deployments from 4G onward. In this context, signal handovers can be inferred by significant signal quality level drops owing to ping-pong reattachment among threshold-based controllers which belong to different neighboring cells. It can be deduced that user experience can be based on non-cell-specific measurements, considering the user experience as a continual measurement along taken trajectory. Handover prediction can thus consist a valid candidate for minimizing service access latency that is critical for V2X applications.

Historic signal quality measurements can be explored and forecast with a two-fold objective. Firstly, the implementation of forecasting software covers what is needed for providing information on the future conditions of the V2I link. Secondly, it implicitly covers the needs for an anomaly detection mechanism by detecting future handovers and general degradation in the quality of experience to provide an early detection mechanism for handovers and increase the resilience of the system for managing a connected vehicle from the scope of the edge infrastructure. A user may be assigned to an ensemble of DNN models that will accurately forecast future experience for all considered measures of interest across multiple steps forward in time. If the signal quality is predicted to peak, then the ability to further increase speed emerges, this choice being bounded by the road limits (e.g. 60 km/h for urban roads, 120 km/h for highways). If the signal starts to drop (imminent handover) then not increasing speed during the drop period is applied that is not further exacerbating the current experience and granting a stable transition to the new cell during handover. RSRQ is usually not available immediately after a handover, so historic NaN entries are replaced with minimum in-sample RSRQ values.

5.3 Providing Live Journey Directions

Choosing a route usually involves consulting attainable speed according to road limits, travel distance between waypoints and road traffic congestion. Under this workflow, congestion data is gathered live and no history is created, meaning that seasonality analysis is discarded by

this work. One may alternatively collect road traffic congestion at e.g. a daily frequency together with a moving average scheme, to provide a statistical description of the first and second moments of a road segment congestion. In general, a lower mean speed implies more traffic or extraneous conditions, while zero mean speed at a road segment implies road blockage, disaster, extreme traffic, etc. If a route is characterized by high (e.g. > 100%) congestion levels, then alternative routes are preferred over that route.

Additional considerations must be accounted for if a vehicle opts for fastest travel time to destination, which translates into choosing shorter, congestion-free road segments. On the other hand, if a vehicle opts for minimal energy consumption (in liters for fuel engines or watts for electric vehicles), either minimum travel distance or conservative speeds under minimally-congested routes are preferred.

The workflow that was followed firstly consists of road safety middleware that provides the basis for associating the starting journey with live road data. Besides collecting needed information on the journey waypoints, necessary calculations for mean speed and congestion are carried out, while data fusion techniques are applied in the end of the journey in order to present a full personalized overview of the travel itinerary that was followed during the operation of the routing mechanism.

Experiments were carried out on small-scale datasets that contain location and signal quality for a vehicle in urban conditions which may or may not include highways, these locations resampled with a fixed striding pattern that reflects the desired period of invocation for the journey routing and live congestion estimation mechanism. Each location that ends up being included is considered a journey waypoint. Between two waypoints, the travel distance along an available route is segmented into steps that bear travel instructions for the vehicle under guidance by the Google Directions JSON API. Each step provides the basis for road traffic analysis that can be performed both for each step and per pair of waypoints.

5.4 Presentation Layer Semantics

At the presentation layer, various information for the vehicle journey may be consulted and this includes available segmented routes, suggested route under suggested speed index given a user plan (fastest time or best consumption), distance and congestion levels. All available routes and relevant information is shown both on per-waypoint and per-step bases along each available path.

Additionally, driving instructions per congested road segment are shown on a per-step basis to complement the driving/passenger experience.

Current mode of work avoids performing a static travel itinerary planning at the time of departure, since journey routing choices are renewed close to the time of arrival at each journey waypoint, given a fixed strided pattern off the original locations. By doing so, the user gains fine-grained control over taken trajectory by being provided the latest road conditions, leading to an intuitive and repeatable plug-and-play behaviour. For simplicity, it is assumed that the number of routes lies between 1 and 3 inclusive, which forms an RGB-colored route map that is easy to follow and can be used to quickly troubleshoot the travel itinerary mechanism that is implemented. Journey waypoints and per-route road segments bear curated information from the requests to the Google Directions servers and may contain distance, live traffic conditions, route, waypoint and segment indexes. Information on locations that are passed more than once during the course of the same journey must be properly consolidated as in the case of steps belonging to the same route index, overlapping at the presentation layer.

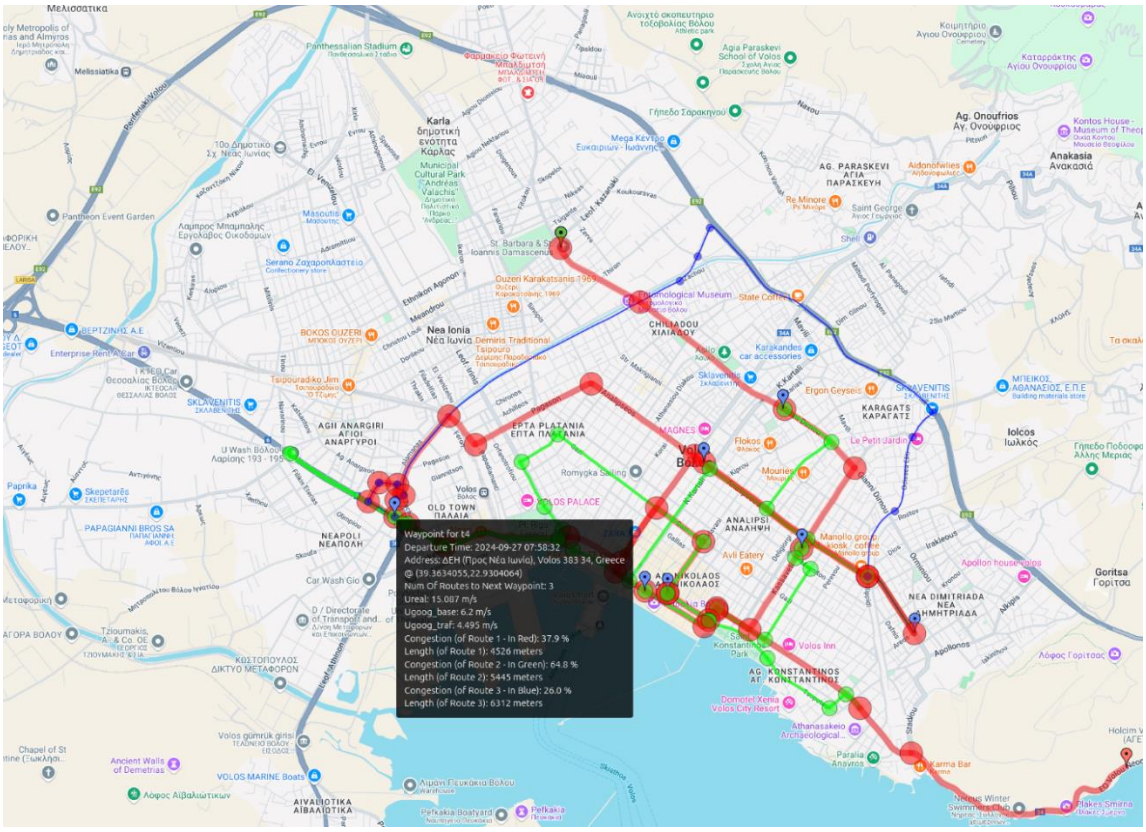

Figure 7. Road Information Presentation based on the Google Maps and Directions API

5.5 Evaluation and Results

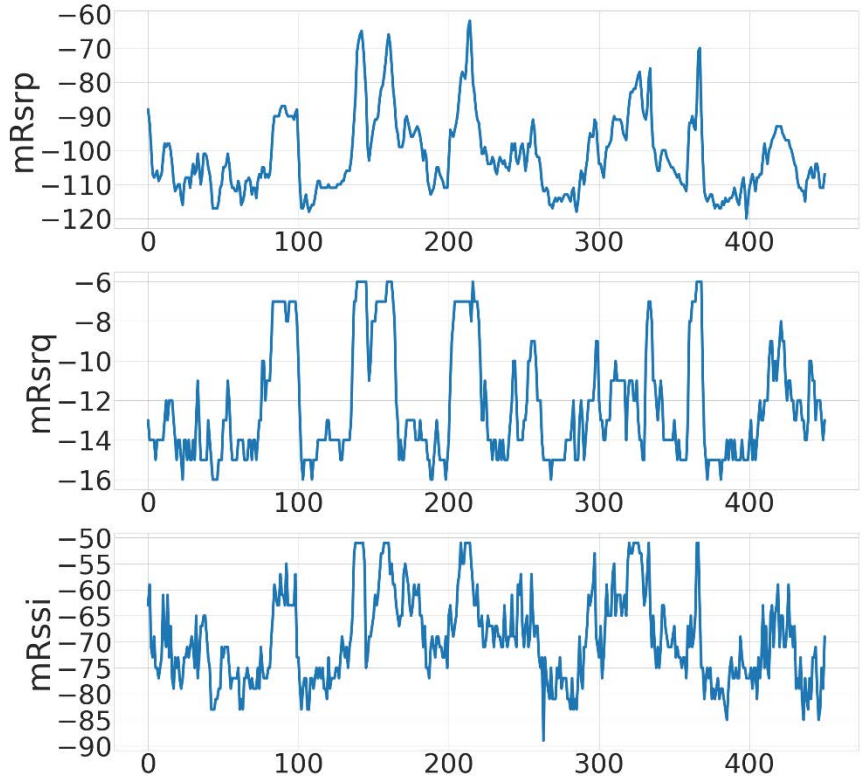

Figure 8. Time Series Visualization for Signal Prediction Dataset

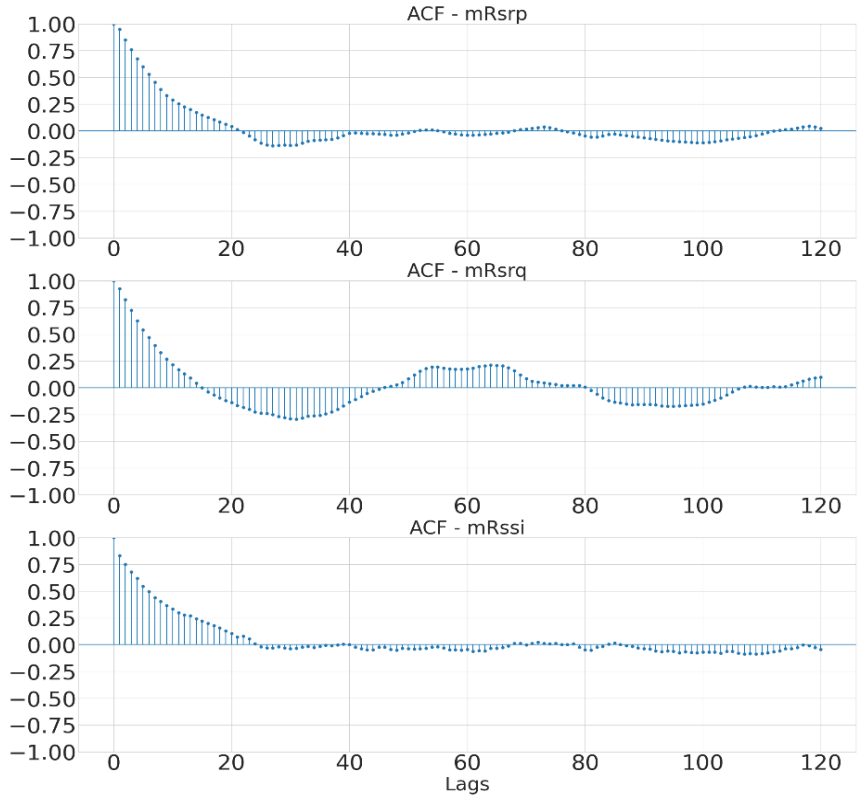

Figure 9. Autocorrelation Plot for Signal Prediction Dataset

Figure 8 visualizes a small urban drive dataset that was collected in the premises of the city of Volos. The QoS metrics considered are:

- RSRP (Reference Signal Received Power) measured in the range of [-140.0, -44.0] dBm
- RSRQ (Reference Signal Received Quality) in the range of [-19.5, -3.0] dB
- RSSI (Received Signal Strength Indicator) with a range of [-120.0, 0.0] dBm

From the autocorrelation plot of Figure 9, it can be deduced that all three metrics contain a medium level of autocorrelation that stretches between lags 20 and 40, hinting on their moderate predictability as multiple time series.

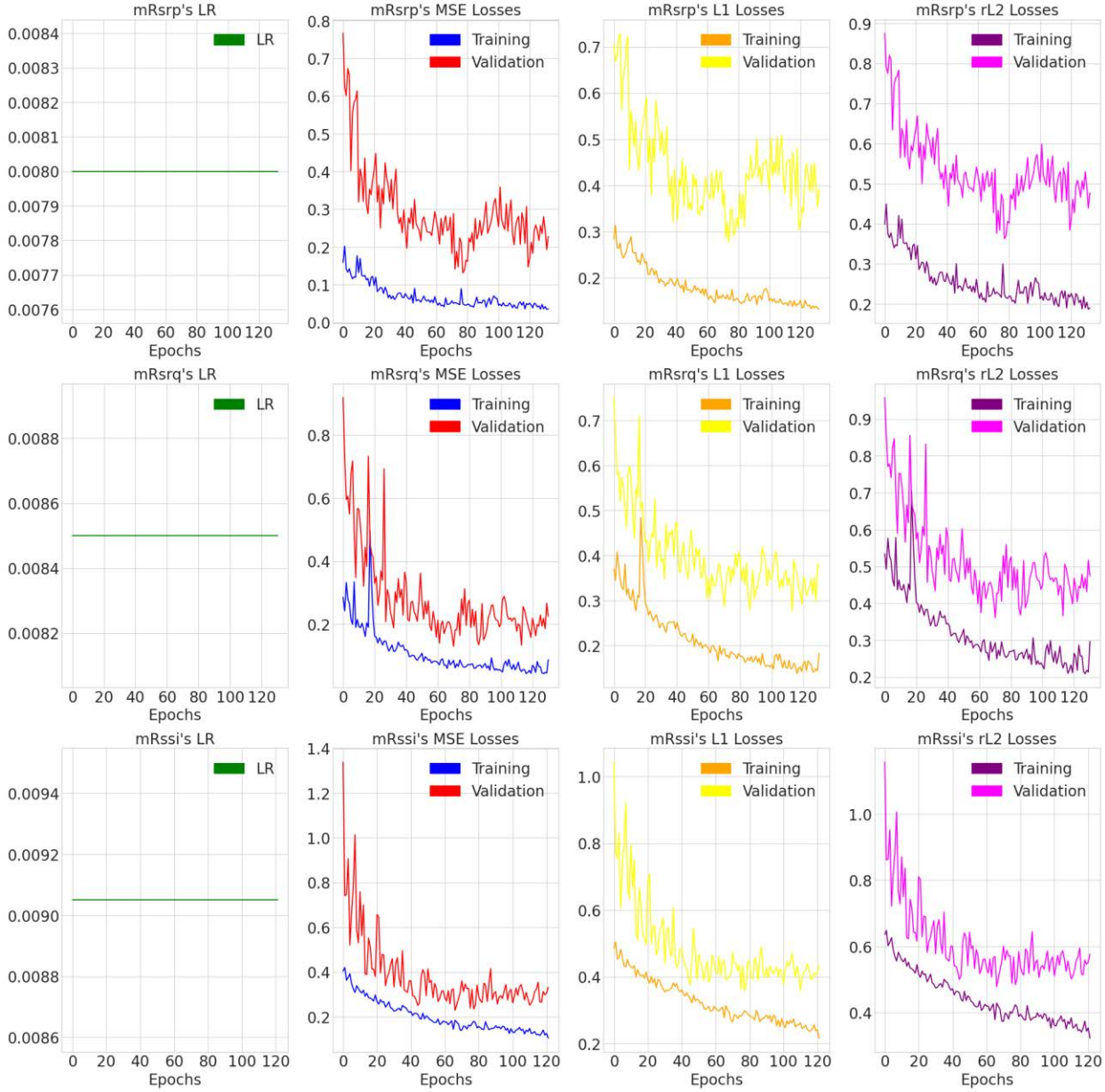

Figure 10. Training and Validation Diagnostics Plots for MSE, MAE, RMSE metrics.

Diagnostic Plots for the training and validation phases (Figure 10) are usually consulted after online training with the chosen models, where of main interest is progressively minimizing the losses after each iteration within a well-defined training budget. A budget of 150 epochs was used for each prediction task. Further, evaluation of model out-of-sample performance was performed by partitioning the user dataset in contiguous train, validation and test sets. The last 50 points belong to the unseen test set, when the previous 50 make up the validation set that is used to evaluate the model weights at each epoch during training. The error in jointly forecasting the QoS measures of for a forecasting task of h steps ahead is quantified by means of Relative Error (RE) and Mean Average Percentage Error (MAPE), where *ih* denotes the *i-th* variable at time *t* + *h* and y represents the actual value at specified index. RE can capture differences in scale and MAPE expresses forecast accuracy as a percentage, while for both metrics lower means better:

$$\mathbf{RE} = \frac{\sqrt{\sum_{i=1}^{K}\sum_{h=1}^{H}\left(y_{ih}-\hat{y}_{ih}\right)^2}}{\sqrt{\sum_{i=1}^{K}\sum_{h=1}^{H}\left(y_{ih}\right)^2}}$$

$$\mathbf{MAPE} = \frac{\sum_{i=1}^{K}\sum_{h=1}^{H}\left|y_{ih}-\hat{y}_{ih}\right|}{\sum_{i=1}^{K}\sum_{h=1}^{H}\left|y_{ih}\right|}100\%$$

The inference phase spans across an arbitrarily large forecast horizon H. A forecasting task of H = 50 with a 6s measurement granularity is defined, allowing for a 5-minute window approximation into user's future performance. The error performances were checked at the t + 1, t + 10 and t + 50 points, while also being accompanied by a plot overview (Figure 11).

The experiment results (Table 1) reveal that per-variable ARIMA model selections are able to provide robust errors across the forecast horizon. Further, a low-order VAR(p) model can be seen to provide errors that grow in a well-behaved manner, while over-parameterized VAR models that originate from the same model set can produce predictions that saturate from early on, our own about the h = 10 point. Fitting the training data with the model once to cover for all desired points in the forecast horizon, favors neither ARIMA-based nor VAR-based models. Both types are severely affected by the repeated bootstrapping that is needed for lengthy forecast horizons and their integration in a windowed retraining scheme is needed if needed. The GRU-based

ensemble model manages to produce state-of-the-art results (Figure 12), essentially generating real-life representative QoS forecasts that can be purposefully used to optimize the user experience.

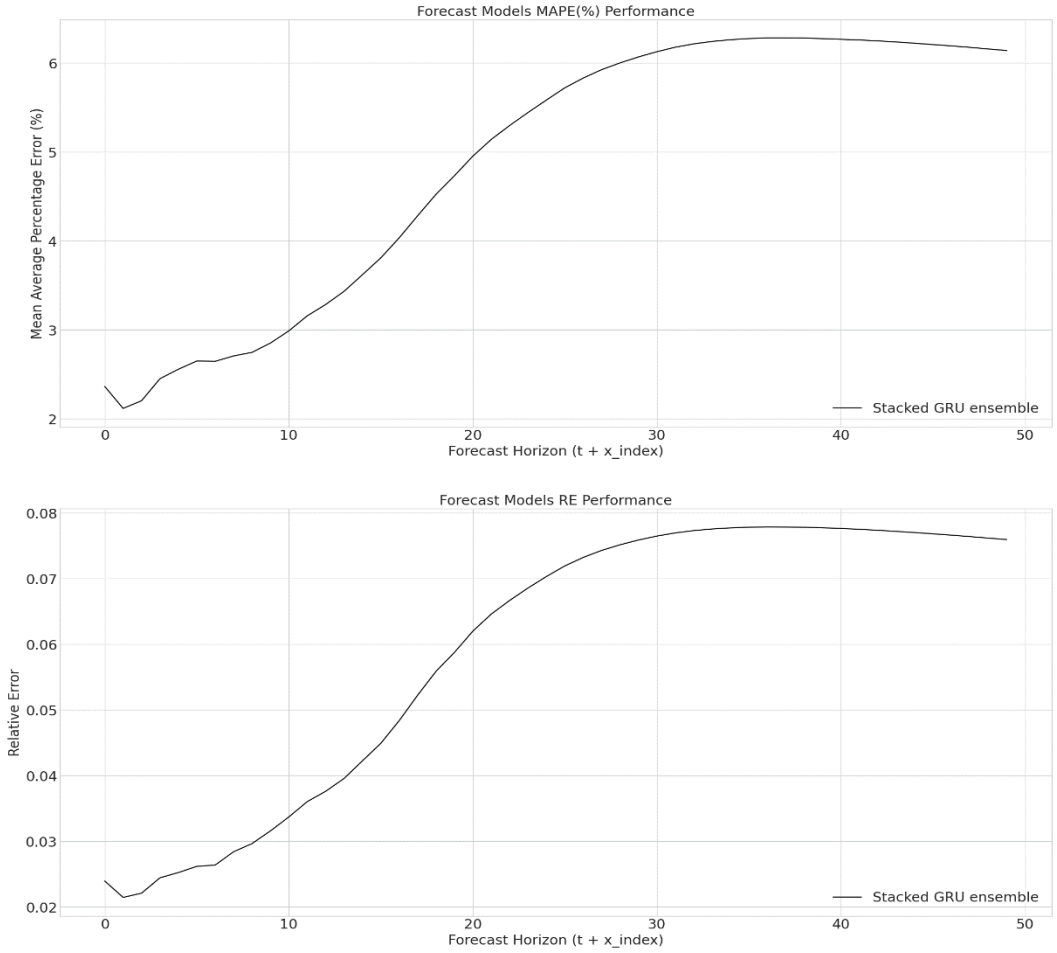

Figure 11. Out-of-sample Forecasting MAPE (%) and RE performance

The models were randomly initialized with a scheme that chooses the max between the fan-out and the fan-in at each model layer. The fan-out refers to the number of outgoing connections and the fan-in refers to the input connections. Additional to the hyperparameters shown in Table 2, the RMSProp optimizer was uniformly used for all target variables. This optimizer controls the oscillations in the variance of the squared gradients by dampening them *more lightly* as the parameter's value reaches 1, permitting higher LRs. Default *rho* values were maintained. The neuron architecture follows a stably descending fashion and is uniform for all models in the ensemble [64,32,16,8] sealed off with a singular dense unit. This particular implementation [14] involves independently training each ensemble model but collaboratively performing the inference phase within the ensemble since forecast points after the first (t+1) are dependent on the intermediate results of the whole ensemble.

Forecast Model	h=1		h=10		h=50	
	RE	MAPE	RE	MAPE	RE	MAPE
ARIMA	0.0404	3.6582	0.0509	4.9978	0.1080	9.6418
VAR(3)	0.0577	6.1140	0.0732	7.3930	0.1143	10.6183
HO-VAR(37)	0.0709	7.5957	0.1204	11.607	0.1187	11.0005
Stacked GRU ensemble	**(0.0337, +-0.0080)**	**(3.2281, +-0.9438)**	**(0.0465, +-0.0078)**	**(3.9116, +-0.5235)**	**(0.0828, +-0.0045)**	**(6.7572, +-0.3930)**

Table 1. Out-of-sample Forecasting MAPE (%) and RE performance

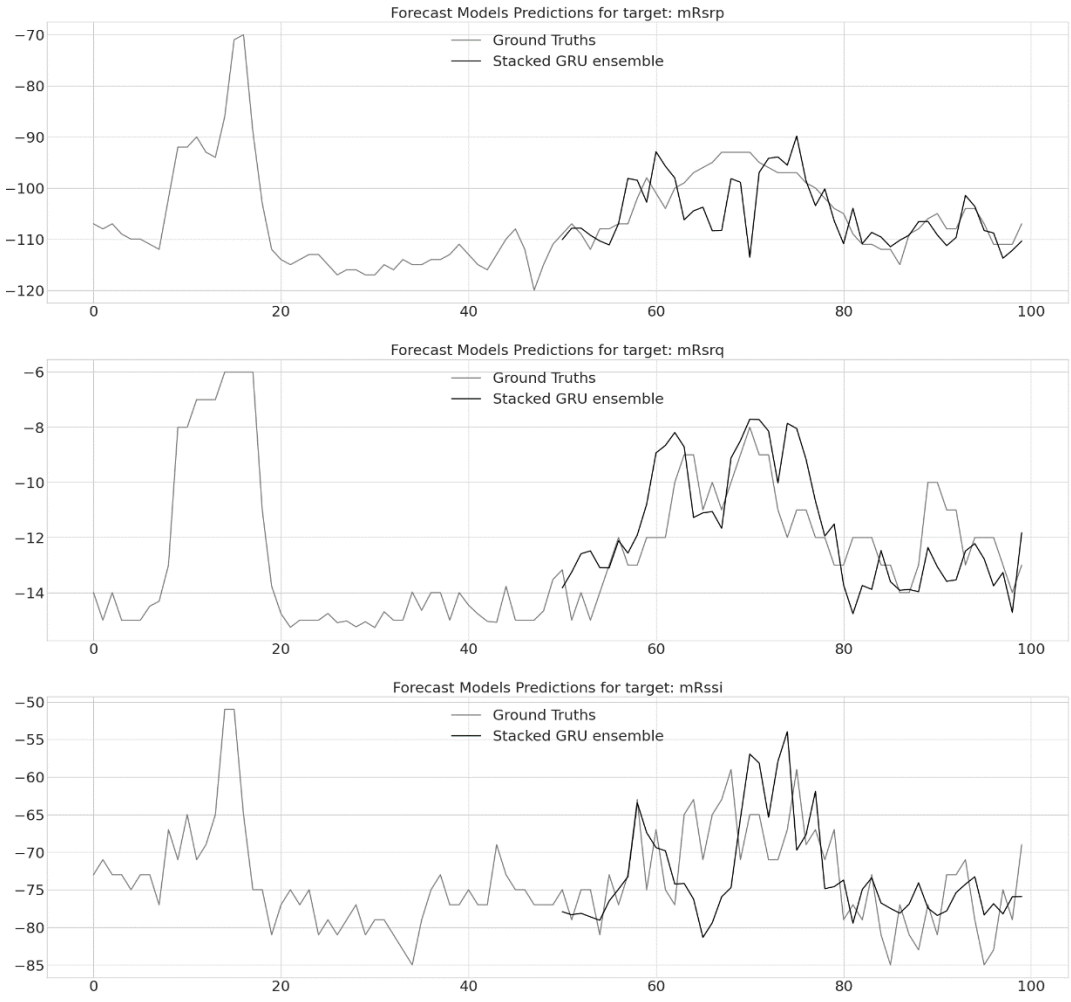

Figure 12. Out-of-sample forecasting realization against ground truths.

	Stacked GRU ensemble
# of Max Training Epochs	(150, 150, 150)
Stopping Patience	(55, 60, 55)
Learning Rate	(9.05e-3, 8.5e-3, 8e-3)
Lagging Depth	(3, 3, 3)
Batch Size	(1, 1, 1)
Parameters (total trainable)	80.088

Table 2. Training Hyperparameters for chosen GRU ensemble model

The specifications of the implementation that was carried out involves the development of Python 3.8 middleware over an Ubuntu 24.04 LTS Host on the side of the road (edge node) that is equipped with a 20-core Intel i7 CPU. The GRU model and relevant components were implemented using Tensorflow 2.4. Providing parallelization abilities to the Python application involved utilizing the multiprocessing Python library that contains implementations for cross-platform process-activating primitives. It is specifically evaluated how each available implementation behaves on scaling the number of users. The traditional fork implementation is denoted with a blue line, while the spawn primitive with a dotted red line. The forkserver process model is finally drawn with a green line.

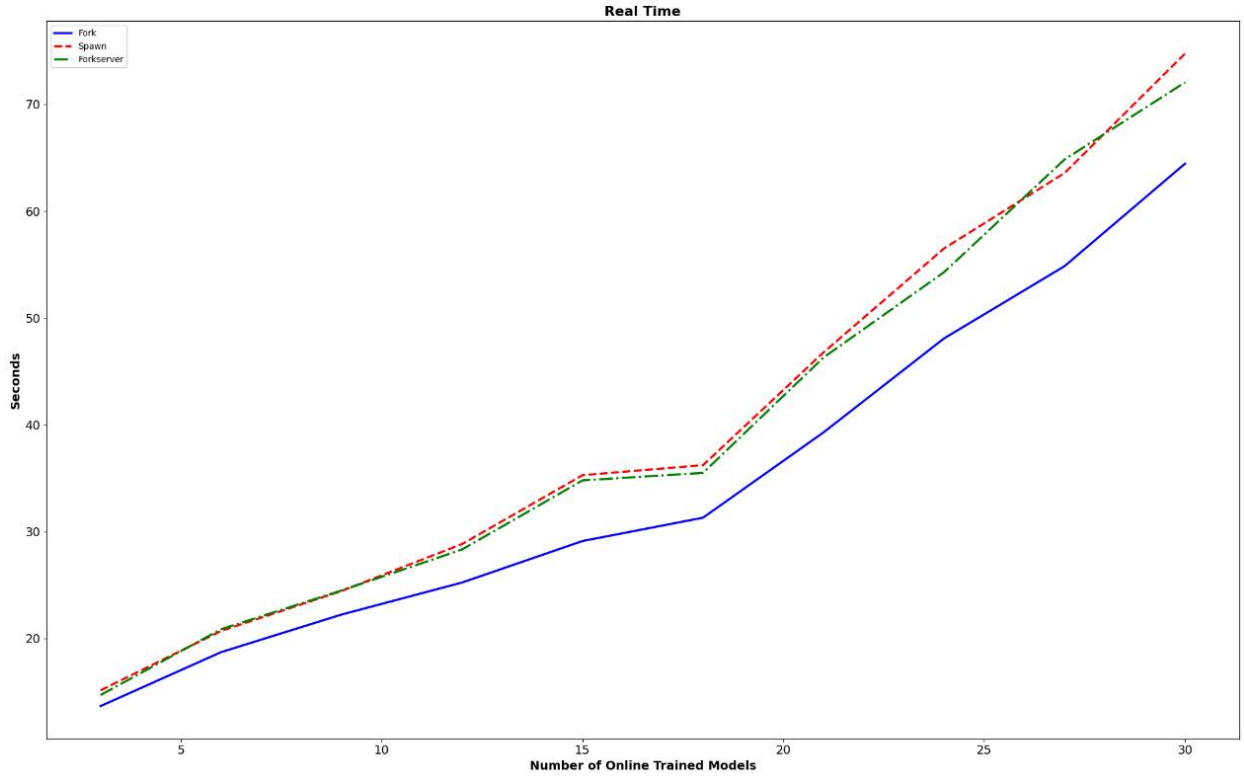

Figure 13. Real time vs Number of Trained Models.

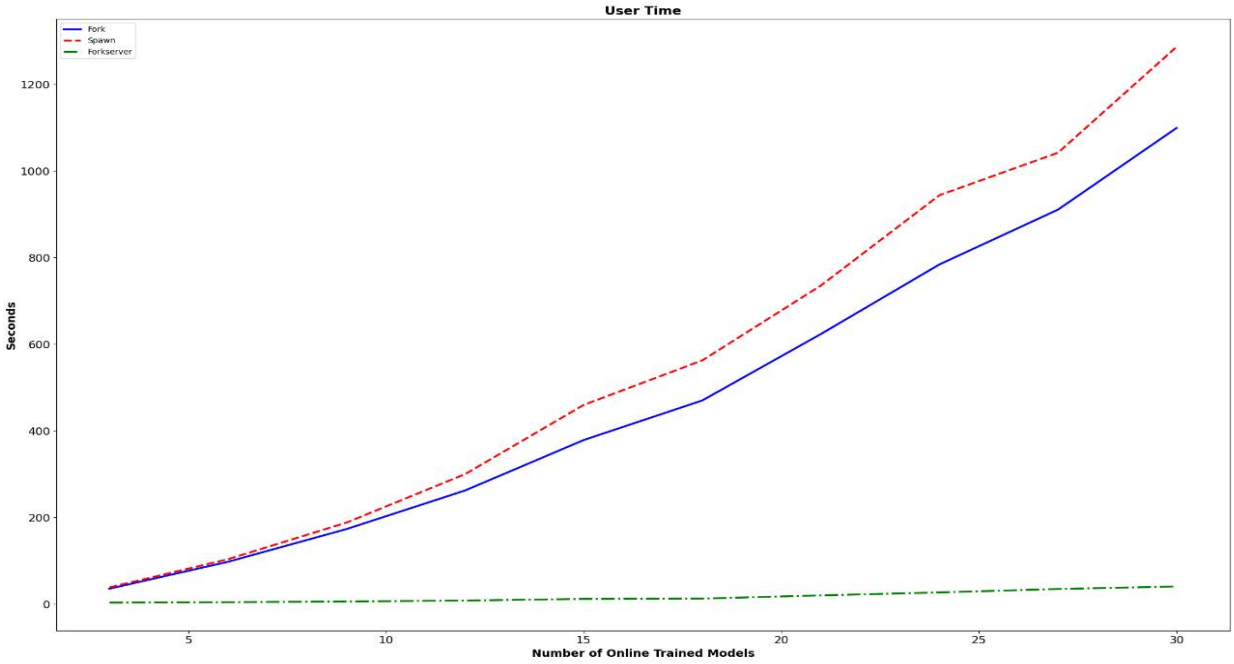

Figure 14. User Time vs Number of Trained Models

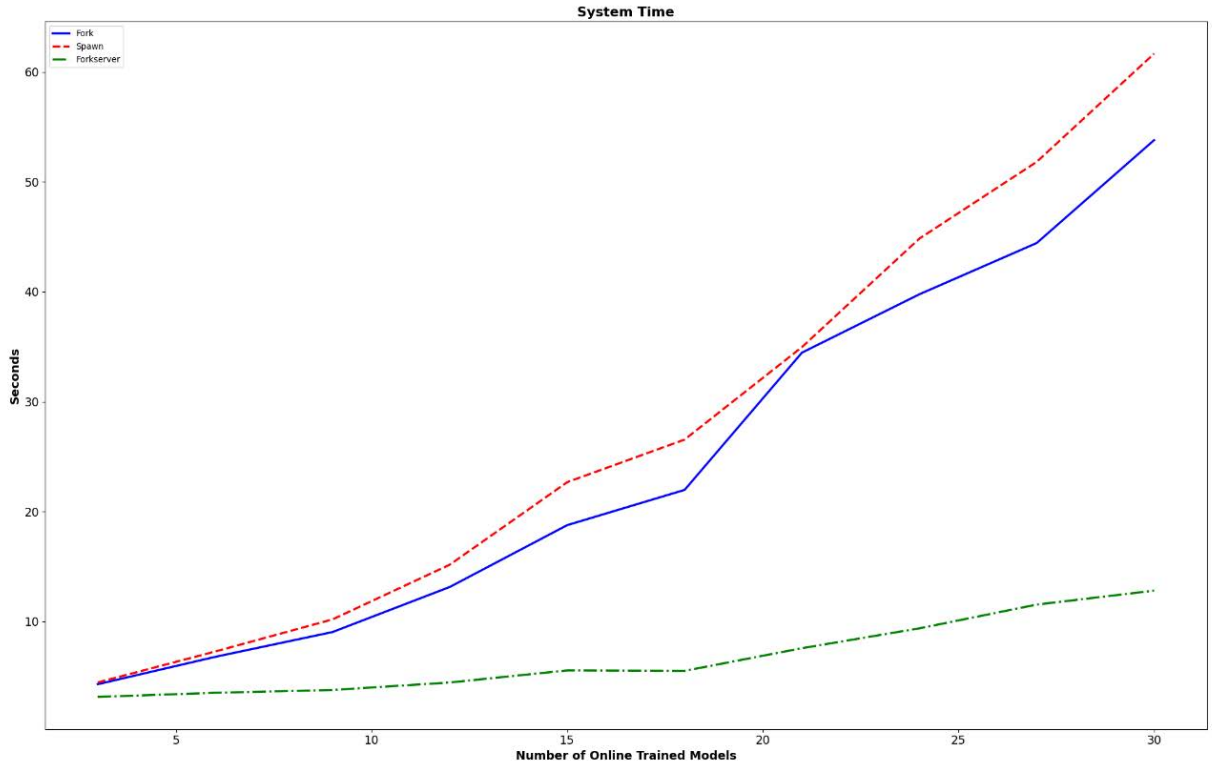

Figure 15. System Time vs Number of Trained Models

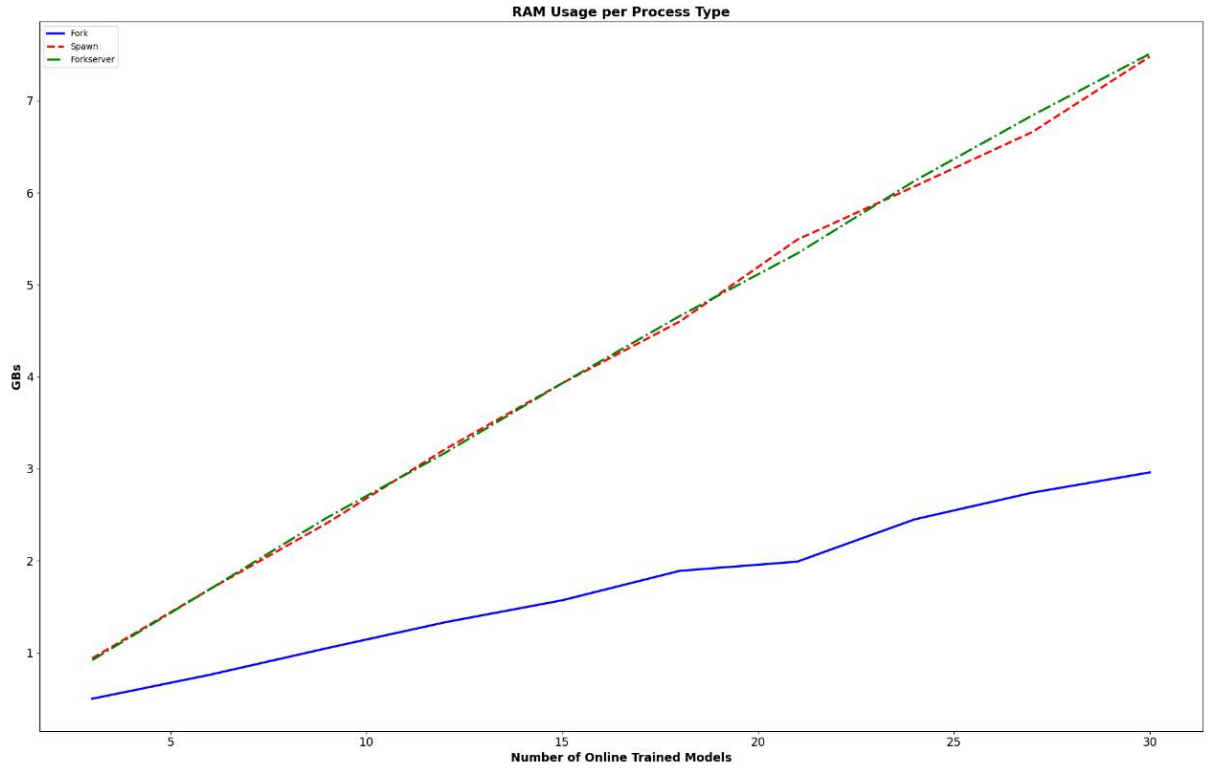

Figure 16. RAM Usage vs Number of Trained Models

The above plots represent multi-user online training and inference of models when the number of users is scaled, thereby for each new user, 3 additional modules are trained to provide them with needed forecasts for each target variable. The amount of memory required to scale such applications in GBs, Real execution time, pure-user level time and Linux system time are investigated. It can be observed that the fork implementation is by far the most economical in terms of memory used (Figure 16), scaling gracefully while being constantly below the other two options. In terms of system (Figure 15) and user (Figure 14) times, the forkserver dominates the other two options, behaving in a detached manner given the increasing number of users. For the actual time of the execution (Figure 13), it can be deduced that the fork option is a clear standout, while spawn and forkserver implementations behave and degrade similarly.

Chapter 6 - Conclusion

This thesis provided an overview of the advancements in modern cloud-enabled wireless networks with a special focus on road safety and V2X. After providing an outline of current research directions and relevant works in literature, this work implemented middleware for supporting road safety applications under live feed of road information that is required to support such applications. Further, by leveraging the GRU model and Deep Learning methods empirically validated with experimentation, this work proved the predictability of multi-metric signal quality time series belonging to network users characterized by mobility with respect to the viability of the underlying model's forecast errors and the number of parameters involved. This workflow is recommended to be followed by applications that require predictions on the V2I link as described in a previous chapter. Chosen methods were evaluated and compared to other available approaches as in peer works but for demanding forecasting horizons concerning smaller datasets to cover for scenarios that would be needed to achieve real-life representative performances under stringent timing. The procedure of selection for DNN models is not trivial and requires implementing precise software with minimized programming errors and inconsistencies to achieve reproducible results under a well-defined budget in e.g. iterations, CPU, memory, etc. There is plenty of room for innovation and automation regarding the integration of intelligent prediction mechanisms in a cloud network environment, with problem areas including efficient scaling, training, validation and inference cost optimization or even more advanced learning schemes such as federated learning, transfer learning, etc.

Future Work

This work presented results that encourage the process of implementing intelligent modeling and forecasting systems for Edge-enabled vehicular networks in terms of signal prediction, resource provisioning and failure detection, them being key to various optimizations towards a higher level of self-organization and in-network automation. The author is interested in studying the effects of dynamically adjusting the training hyperparameters in order to fortify the learning process against problems of model sensitivity to random initialization and the vanishing gradient problem.

References

1. P. Marsch, O. Bulakci, O. Queseth, M. Boldi, “5G System Design”, 2018, Wiley
2. H. Hartenstein and L. Laberteaux, “A tutorial survey on vehicular ad hoc networks”, IEEE Communications Magazine, vol. 46, no. 6, pp.164–171, Jun. 2008.
3. K. Alexandris, N. Nikaein, R. Knopp, C. Bonnet, “Analyzing X2 Handover in LTE/LTE-A”, May 2016
4. C. Y. Chang, K. Alexandris, N. Nikaein, K. Katsalis, T. Spyropoulos (2016), MEC architectural implications for LTE/LTE-A networks. In Proceedings of the Workshop on Mobility in the Evolving Internet Architecture (MobiArch '16). Association for Computing Machinery, New York, NY, USA, 13–18. https://doi.org/10.1145/2980137.2980139
5. ETSI Group Specification etsi.org/deliver/etsi_gs/MEC/001_099/012/02.01.01_60/gs_MEC012v020101p.pdf
6. F. Giust, V. Sciancalepore, D. Sabella, M. C. Filippou, S. Mangiante, W. Featherstone, D. Munaretto, "Multi-Access Edge Computing: The Driver Behind the Wheel of 5G-Connected Cars," in *IEEE Communications Standards Magazine*, vol. 2, no. 3, pp. 66-73, SEPTEMBER 2018, doi: 10.1109/MCOMSTD.2018.1800013
7. D. Bega, M. Gramaglia, R. Perez, M. Fiore, A. Banchs and X. Costa-Pérez, "AI-Based Autonomous Control, Management, and Orchestration in 5G: From Standards to Algorithms," in IEEE Network, vol. 34, no. 6, pp. 14-20, November/December 2020, doi: 10.1109/MNET.001.2000047
8. A. Boudi, M. Bagaa, P. Pöyhönen, T. Taleb and H. Flinck, "AI-Based Resource Management in Beyond 5G Cloud Native Environment," in *IEEE Network*, vol. 35, no. 2, pp. 128-135, March/April 2021
9. G. White, A. Palade and S. Clarke, "Forecasting QoS Attributes Using LSTM Networks," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489052
10. A. Kousaridas, R. P. Manjunath, J. Perdomo, C. Zhou, E. Zielinski, S. Schmitz, A. Pfadler, "QoS Prediction for 5G Connected and Automated Driving," in *IEEE Communications Magazine*, vol. 59, no. 9, pp. 58-64, September 2021, doi: 10.1109/MCOM.110.2100042
11. H. Yin, X. Guo, P. Liu, X. Hei, Y. Gao, "Predicting Channel Quality Indicators for 5G Downlink Scheduling in a Deep Learning Approach", https://arxiv.org/abs/2008.01000
12. F. Tang, Y. Kawamoto, N. Kato and J. Liu, "Future Intelligent and Secure Vehicular Network Toward 6G: Machine-Learning Approaches," in *Proceedings of the IEEE*, vol. 108, no. 2, pp. 292-307, Feb. 2020, doi: 10.1109/JPROC.2019.2954595
13. H. Woo, M. Sugimoto, J. Wu, Y. Tamura, A. Yamashita, H. Asama, (2018). Trajectory Prediction of Surrounding Vehicles Using LSTM Network
14. https://github.com/christosnanis/master-thesis

Abbreviations

Term	Meaning
V2X	Vehicle-to-Everything
QoS	Quality of Service
5G	5th Generation
B5G	Beyond 5th Generation
4G	4th Generation
LTE	Long-Term Evolution
MSE	Mean Squared Error
MAE	Mean Average Error
RMSE	Root Mean Squared Error
MAPE	Mean Average Percentage Error
RE	Relative Error
GRU	Gated Recurrent Unit
GSM	Global System for Mobile communications
Kbit	Kilobit
SMS	Short Message Service
GHz	Gigahertz
NAS	Non-Access Stratum
RRC	Radio Resource Control
PDCP	Packet Data Convergence Protocol
RLC	Radio Link Control
3GPP	3rd Generation Partnership Project
WiFi	Wireless Fidelity
eMBB	Enhanced Mobile Broadband
uRLLC	Ultra Low Latency Communications
mmTC	Massive Machine-Type Communications
GPS	Global Position Satellite
LIDAR	Light Detection and Ranging
MIMO	Multiple Input Multiple Output
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
ITS	Intelligent Transportation Systems
UE	User Equipment
SLA	Service-Level Agreement
RAT	Radio Access Technology
RSU	Road Side Unit
MEC	Multi-access Edge Computing
NFV	Network Function Virtualization
SDN	Software Defined Network

RAN	Radio Access Network
ML	Machine Learning
AI	Artificial Intelligence
IoT	Internet of Things
NB	Narrowband
PCF	Policy Control Function
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
ARIMA	Autoregressive Integrated Moving Average
VNF	Virtual Network Function
NR	New Radio
CQI	Channel Quality Indication
dB	decibels
VAR	Vector Autoregressive
NaN	Not a Number