AI-Enhanced Methods in Autonomous Systems: Large Language Models, DL Techniques, and Optimization Algorithms

November 2023

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Science in Computer Science
by Irene de Zarzà i Cubero

Supervisor: Dr. Carlos Miguel Tavares de Araujo Cesariny Calafate

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

"What is now proved was once only imagined." - W. Blake

Acknowledgements

As I reflect upon the journey that culminated in this thesis, I am overwhelmed with gratitude for the individuals who have guided, supported, and inspired me throughout.

First and foremost, my deepest appreciation goes to my advisor, Carlos T. Calafate. Your wisdom, patience, and unwavering belief in my abilities have been the guiding light at every challenging turn, pushing me to scale greater heights. To Gemma Roig, your insightful feedback and profound expertise were instrumental in shaping many sections of this work. The time you invested in our discussions and brainstorming sessions have been invaluable.

I would like to express my sincere gratitude to the incredible community at UPV. Special mentions are in order for my co-authors Pietro Manzoni, Juan Carlos Cano and Enrique Hernández, whose collaborations have enriched this research immensely.

The enriching environment and relentless pursuit of knowledge at GOETHE-University have been a cornerstone of my academic journey. I am especially thankful to everyone from the Group Computational Vision and AI, and the Center for Cognition and Computation, who provided constructive feedback and constant encouragement.

My time with the logistics group at TU Darmstadt was enlightening, and I cherish the intellectual discussions and shared insights that influenced this thesis. The team at the Centre for Intelligent Multidimensional Data Analysis (HK Science Park), the Mathematical Institute at the University of Oxford and the EE Department at City University of Hong Kong offered me a diverse perspective that I am immensely grateful for.

On a personal note, J. de Curtò, my dear husband, your unwavering support and love have been my anchor. Your faith in me, even during moments of self-doubt, was a constant reassurance that kept me going. Your sacrifices and belief in me have been the bedrock on which this achievement stands. To my family, whose love and encouragement know no bounds.

In conclusion, while these words are but a humble attempt to convey my gratitude, the deep sense of indebtedness I feel can only truly be understood in the silent moments of reflection. To all who walked this path with me, be it in significant strides or subtle steps, I am eternally thankful.

Abstract

The proliferation of autonomous systems, and their increasing integration with day-to-day human life, have opened new frontiers of research and development. Within this scope, the current thesis dives into the multifaceted applications of Large Language Models (LLMs), Deep Learning (DL) techniques, and Optimization Algorithms within the realm of these autonomous systems. Drawing from the principles of AI-enhanced methods, the studies encapsulated within this work converge on the exploration and enhancement of different autonomous systems ranging from B5G Truck Platooning Systems, Multi-Agent Systems (MASs), Unmanned Aerial Vehicles, Forest Fire Area Estimation, to the early detection of diseases like Glaucoma.

A key research focus, pursued in this work, revolves around the innovative deployment of adaptive PID controllers in vehicle platooning, facilitated through the integration of LLMs. These PID controllers, when infused with AI capabilities, offer new possibilities in terms of efficiency, reliability, and security of platooning systems. We developed a DL model that emulates an adaptive PID controller, thereby showcasing its potential in AI-enabled radio and networks. Simultaneously, our exploration extends to multi-agent systems, proposing an Extended Coevolutionary (EC) Theory that amalgamates elements of coevolutionary dynamics, adaptive learning, and LLM-based strategy recommendations. This allows for a more nuanced and dynamic understanding of the strategic interactions among heterogeneous agents in MASs.

Moreover, we delve into the realm of Unmanned Aerial Vehicles (UAVs), proposing a system for video understanding that employs a language-based world-state history of events and objects present in a scene captured by a UAV. The use of LLMs here enables open-ended reasoning such as event forecasting with minimal human intervention. Furthermore, an alternative DL methodology is applied for the estimation of the affected area during forest fires. This approach leverages a novel architecture called TabNet, integrated with Transformers, thus providing accurate and efficient area estimation.

In the field of healthcare, our research outlines a successful early detection methodology for glaucoma. Using a three-stage training approach with EfficientNet on retinal images, we achieved high accuracy in detecting early signs of this disease.

Across these diverse applications, the core focus remains: the exploration of advanced AI methodologies within autonomous systems. The studies within this thesis seek to demonstrate the power and potential of AI-enhanced techniques in tackling complex problems within these systems. These in-depth investigations, experimental analyses, and developed solutions shed light on the transformative potential of AI methodologies in improving the efficiency, reliability, and security of autonomous systems, ultimately contributing to future research and development in this expansive field.

Resumen

La proliferación de sistemas autónomos y su creciente integración en la vida humana cotidiana han abierto nuevas fronteras de investigación y desarrollo. Dentro de este ámbito, la presente tesis se adentra en las aplicaciones multifacéticas de los LLMs (Large Language Models), técnicas de DL (Deep Learning) y algoritmos de optimización en el ámbito de estos sistemas autónomos. A partir de los principios de los métodos potenciados por la Inteligencia Artificial (IA), los estudios englobados en este trabajo convergen en la exploración y mejora de distintos sistemas autónomos que van desde sistemas de platooning de camiones en sistemas de comunicaciones Beyond 5G (B5G), Sistemas Multi-Agente (SMA), Vehículos Aéreos No Tripulados (UAV), estimación del área de incendios forestales, hasta la detección temprana de enfermedades como el glaucoma.

Un enfoque de investigación clave, perseguido en este trabajo, gira en torno a la implementación innovadora de controladores PID adaptativos en el platooning de vehículos, facilitada a través de la integración de los LLMs. Estos controladores PID, cuando se infunden con capacidades de IA, ofrecen nuevas posibilidades en términos de eficiencia, fiabilidad y seguridad de los sistemas de platooning. Desarrollamos un modelo de DL que emula un controlador PID adaptativo, mostrando así su potencial en las redes y radios habilitadas para IA. Simultáneamente, nuestra exploración se extiende a los sistemas multiagente, proponiendo una Teoría Coevolutiva Extendida (TCE) que amalgama elementos de la dinámica coevolutiva, el aprendizaje adaptativo y las recomendaciones de estrategias basadas en LLMs. Esto permite una comprensión más

matizada y dinámica de las interacciones estratégicas entre agentes heterogéneos en los SMA.

Además, nos adentramos en el ámbito de los vehículos aéreos no tripulados (UAVs), proponiendo un sistema para la comprensión de vídeos que crea una log de la historia basada en la descripción semántica de eventos y objetos presentes en una escena capturada por un UAV. El uso de los LLMs aquí permite razonamientos complejos como la predicción de eventos con mínima intervención humana. Además, se aplica una metodología alternativa de DL para la estimación del área afectada durante los incendios forestales. Este enfoque aprovecha una nueva arquitectura llamada TabNet, integrada con Transformers, proporcionando así una estimación precisa y eficiente del área.

En el campo de la salud, nuestra investigación esboza una metodología exitosa de detección temprana del glaucoma. Utilizando un enfoque de entrenamiento de tres etapas con EfficientNet en imágenes de retina, logramos una alta precisión en la detección de los primeros signos de esta enfermedad.

A través de estas diversas aplicaciones, el foco central sigue siendo la exploración de metodologías avanzadas de IA dentro de los sistemas autónomos. Los estudios dentro de esta tesis buscan demostrar el poder y el potencial de las técnicas potenciadas por la IA para abordar problemas complejos dentro de estos sistemas. Estas investigaciones en profundidad, análisis experimentales y soluciones desarrolladas arrojan luz sobre el potencial transformador de las metodologías de IA en la mejora de la eficiencia, fiabilidad y seguridad de los sistemas autónomos, contribuyendo en última instancia a la futura investigación y desarrollo en este amplio campo.

Resum

La proliferació de sistemes autònoms i la seua creixent integració en la vida humana quotidiana han obert noves fronteres de recerca i desenvolupament. Dins d'aquest àmbit, la present tesi s'endinsa en les aplicacions multifacètiques dels LLMs (Large Language Models), tècniques de DL (Deep Learning) i algoritmes d'optimització en l'àmbit d'aquests sistemes autònoms. A partir dels principis dels mètodes potenciats per la Intel·ligència Artificial (IA), els estudis englobats en aquest treball convergeixen en l'exploració i millora de diferents sistemes autònoms que van des de sistemes de platooning de camions en sistemes de comunicacions Beyond 5G (B5G), Sistemes Multi-Agent (SMA), Vehicles Aeris No Tripulats (UAV), estimació de l'àrea d'incendis forestals, fins a la detecció precoç de malalties com el glaucoma.

Un enfocament de recerca clau, perseguit en aquest treball, gira entorn de la implementació innovadora de controladors PID adaptatius en el platooning de vehicles, facilitada a través de la integració dels LLMs. Aquests controladors PID, quan s'infonen amb capacitats d'IA, ofereixen noves possibilitats en termes d'eficiència, fiabilitat i seguretat dels sistemes de platooning. Desenvolupem un model de DL que emula un controlador PID adaptatiu, mostrant així el seu potencial en les xarxes i ràdios habilitades per a IA. Simultàniament, la nostra exploració s'estén als sistemes multi-agent, proposant una Teoria Coevolutiva Estesa (TCE) que amalgama elements de la dinàmica coevolutiva, l'aprenentatge adaptatiu i les recomanacions d'estratègies basades en LLMs. Això permet una comprensió més matissada i dinàmica de les interaccions es-

tratègiques entre agents heterogenis en els SMA.

A més, ens endinsem en l'àmbit dels Vehicles Aeris No Tripulats (UAVs), proposant un sistema per a la comprensió de vídeos que crea un registre de la història basat en la descripció semàntica d'esdeveniments i objectes presents en una escena capturada per un UAV. L'ús dels LLMs aquí permet raonaments complexos com la predicció d'esdeveniments amb mínima intervenció humana. A més, s'aplica una metodologia alternativa de DL per a l'estimació de l'àrea afectada durant els incendis forestals. Aquest enfocament aprofita una nova arquitectura anomenada TabNet, integrada amb Transformers, proporcionant així una estimació precisa i eficient de l'àrea.

En el camp de la salut, la nostra recerca esbossa una metodologia exitosa de detecció precoç del glaucoma. Utilitzant un enfocament d'entrenament de tres etapes amb EfficientNet en imatges de retina, aconseguim una alta precisió en la detecció dels primers signes d'aquesta malaltia.

A través d'aquestes diverses aplicacions, el focus central continua sent l'exploració de metodologies avançades d'IA dins dels sistemes autònoms. Els estudis dins d'aquesta tesi busquen demostrar el poder i el potencial de les tècniques potenciades per la IA per a abordar problemes complexos dins d'aquests sistemes. Aquestes investigacions en profunditat, anàlisis experimentals i solucions desenvolupades llançen llum sobre el potencial transformador de les metodologies d'IA en la millora de l'eficiència, fiabilitat i seguretat dels sistemes autònoms, contribuint en última instància a la futura recerca i desenvolupament en aquest ampli camp.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Autonomous systems have metamorphosed into one of the cornerstones of modern technological evolution, fundamentally altering our perspective on interconnectedness, efficiency, and decision-making [1, 2, 3]. As these systems become increasingly embedded in our day-to-day experiences [4] – from vehicle coordination on busy highways to unmanned aerial oversight and healthcare diagnostics – the imperative to augment their capabilities grows ever more pressing. A pivotal facet of this augmentation is the seamless integration of advanced Artificial Intelligence (AI) methodologies [5, 6, 7]. This doctoral thesis, titled "AI-Enhanced Methods in Autonomous Systems: Large Language Models, DL Techniques, and Optimization Algorithms," undertakes an in-depth exploration of such methods, rendering insights and innovations that could redefine the landscape of autonomous systems.

Central to our exploration is the intersection of Large Language Models (LLMs) with Deep Learning (DL) techniques and Optimization Algorithms [8]. This convergence facilitates groundbreaking innovations in autonomous systems, encompassing a gamut of applications – from B5G Truck Platooning Systems [9] and Multi-Agent Systems [10, 11], to Unmanned Aerial Vehicles [12] and early disease detection methodologies [13, 14, 15].

Vehicle Platooning and Network Communications: In the exhilarating domain of vehicle platooning, especially in the context of Beyond 5G (B5G) networks, a paradigm shift is afoot. Vehicles coordinated in platoons, often traveling at high speeds with minimal spacing, necessitate an impeccable degree of synchronization and reliability [16]. This work reveals that the utilization of LLMs, like the GPT-3.5-turbo, in conjunction with adaptive PID controllers [17, 18, 19], can profoundly augment the safety and efficiency of platooning systems, particularly when navigating the challenges posed by communication latencies, packet losses, and varied communication ranges.

Strategic Interactions in Multi-Agent Systems: Venturing into the intricate world of Multi-Agent Systems (MASs), traditional game theory models, albeit instrumental, often grapple with capturing the dynamism and heterogeneity inherent to modern MASs. Herein, we introduce an Extended Coevolutionary (EC) Theory, enriched with LLM-based strategic insights [20]. This novel framework recognizes the multifaceted interactions, diverse risk appetites, and learning capabilities of agents, offering a robust model for emergent cooperative behavior, even amidst disruptions.

Unmanned Aerial Vehicles and Scene Comprehension: A prominent advancement, worth highlighting, is our work with Unmanned Aerial Vehicles (UAVs). By integrating Large Language Models [21, 22], such as BLIP-2 [23, 24] and GPT-3 [25, 26, 27], we lay the groundwork for an intelligent video understanding system. This system translates the visual world of UAVs into a language-based narrative, providing comprehensive descriptions, actionable insights, and predictive forecasts.

Forest Fire Area Estimation: On the environmental front, the devastating impact of forest fires underscores the need for precise area estimation. Leveraging the prowess of the transformer-based architecture TabNet [28], this research demonstrates a state-of-the-art methodology for predicting affected areas, thereby enabling informed disaster management decisions.

Early Disease Detection: Lastly, in the realm of healthcare, early detection often makes the difference between manageable treatment and dire consequences. The research presents an optimized three-stage training approach using EfficientNet [29, 30, 31], showcasing remarkable accuracy in the early detection of Glaucoma from retinal images [32].

As we navigate through this thesis, the mathematical intricacies and methodologies underlying each application are elucidated, providing a robust foundation for understanding the transformative potential of AI-enhanced techniques. Through experimental analyses and developed solutions, this work aims to illustrate not just the contemporary significance, but also the futuristic vision of

AI methodologies in autonomous systems. In essence, this research endeavors to accentuate the synergistic potential of LLMs, DL techniques, and optimization algorithms in reimagining the capabilities of autonomous systems [33, 34], thereby lighting the path for subsequent researchers and developers in this ever-evolving field.

1.1 Objectives

The main objective of this thesis is to enhance the capabilities of autonomous systems through the integration of advanced AI methodologies, particularly focusing on the application of LLMs, DL techniques, and Optimization Algorithms. To achieve this goal, the following specific objectives have been tackled:

- To explore the integration of LLMs within the control loop of autonomous vehicle platoons, enhancing communication and decision-making processes in B5G networks.
- To develop an Extended Coevolutionary (EC) framework that incorporates LLM-based strategy recommendations, fostering cooperative behavior in MASs.
- To design and implement an intelligent semantic scene understanding system for UAVs using a pipeline of LLMs.
- To employ advanced ML and DL techniques, particularly TabNet, for accurate estimation of burned areas in forest fires, aiding in environmental management and disaster response.
- To innovate an optimized three-stage training procedure for the early detection of Glaucoma using EfficientNet variants, demonstrating high accuracy and resource efficiency.

These objectives serve as milestones towards the realization of the thesis' main goal. Each chapter of this dissertation is dedicated to addressing these objectives, with comprehensive experimental analyses to validate the proposed methodologies and solutions. As we navigate through the mathematical intricacies and practical applications of each proposed method, we aim to demonstrate the transformative potential of AI-enhanced techniques in autonomous systems.

1.2 Methodology

The methodological approach of this thesis is rooted in a systematic exploration and development of AI-enhanced methods within autonomous systems. A rigorous, multi-stage strategy was employed to address the research questions posed by the objectives. The following presents the structured methodology undertaken:

1. **Literature Review:** A comprehensive review of existing literature was conducted to identify the state-of-the-art in AI methodologies applied to autonomous systems, including the latest advancements in LLMs, DL techniques, and optimization algorithms.

2. **Problem Formulation:** Based on the literature review, specific challenges within the domain of autonomous systems were identified, and pertinent research questions were formulated. These questions guided the subsequent stages of research.

3. **System Design and Modeling:** For each objective, appropriate system models were developed. This included the design of control loops for vehicle platooning, EC frameworks for MASs, semantic scene understanding systems for UAVs, fire area estimation models, and diagnostic models for early disease detection.

4. **Algorithm Development:** Tailored algorithms leveraging LLMs, DL, and optimization were conceptualized and developed to solve the formulated problems. This stage involved iterative prototyping, testing, and refinement.

5. **Experimental Setup and Data Collection:** Experiments were designed to rigorously test the developed models and algorithms. This included setting up simulations, collecting datasets, and ensuring robust experimental protocols.

6. **Evaluation and Analysis:** The performance of the proposed solutions was evaluated using a variety of metrics appropriate to each domain. Results were analyzed to draw conclusions about the efficacy and efficiency of the AI-enhanced methods.

7. **Validation and Verification:** Where possible, solutions were validated using real-world data and scenarios to verify the practical applicability of the proposed methods.

8. **Documentation and Dissemination:** Findings were meticulously documented, ensuring that the research process and outcomes were transparent and reproducible. Results were also disseminated through publications and presentations to the scientific community.

This structured approach ensured a comprehensive investigation into each research question, providing a robust framework for the development and evaluation of AI-enhanced methods in autonomous systems. Each chapter of this thesis corresponds to a distinct phase in this methodology, collectively contributing to the field's advancement and setting the groundwork for future research endeavors.

In the problem formulation phase, the research was steered by several questions that emerged as a result of the initial literature survey. These questions were meticulously crafted to dissect the core challenges and opportunities within the autonomous systems landscape. Key among them were: How can LLMs like GPT-3.5-turbo be effectively integrated into the control systems of autonomous vehicle platoons to enhance communication and decision-making? What are the potential dynamics of cooperation and defection in MASs when influenced by LLM-based strategic recommendations within an EC framework? Can the integration of LLMs and Visual Language Models (VLMs) in UAVs lead to a breakthrough in semantic scene understanding that transcends current limitations? What DL architectures and models most effectively predict the extent of burned areas in forest fires, and how can these predictions be optimized to assist in disaster management? Lastly, how can the efficiency and accuracy of disease detection, specifically glaucoma, be maximized through advanced neural networks like EfficientNet? Addressing these questions constituted the foundation of our research, ensuring that the developed methodologies not only addressed theoretical gaps but were also attuned to practical, real-world applications and implications. Each chapter thus not only pursues these questions with rigor but also contributes to the body of knowledge that shapes the field of autonomous systems.

Chapter 2

Background

In an age marked by the exponential evolution of AI, our understanding and approach to autonomous systems have been fundamentally altered. Autonomous systems, operating independently of human intervention, have steadily embedded themselves in our daily lives. From transportation networks to smart homes, autonomous robotics to healthcare, they have drastically expanded the scope and capacity of technological integration. At the heart of this revolution are the interwoven threads of LLMs, DL techniques, and Optimization Algorithms, all of which have fueled and catalyzed these developments.

LLMs have emerged as powerful tools for tasks involving natural language understanding, information retrieval, and knowledge extraction. Their ability to comprehend, generate, and interact in human language, and to produce coherent, contextually accurate text, has made them valuable in various applications. These models, including OpenAI's GPT series, excel at an array of tasks from text generation, translation, to answering queries, among others. Their ability to incorporate and interpret vast swaths of information has unlocked unique potential in developing dynamic and adaptive systems, thereby making them pivotal in the autonomous systems discussed in this work.

Beyond Language Models (BLIP-2) [23, 24], a significant evolution of LLMs, delve further into the intersection of Machine Learning (ML) and human-like cognitive abilities. Expanding upon its predecessors, the BLIP-2 model offers enhancements such as multi-modal learning, better representation of learned information, and increased language understanding capacity. Its underlying design and functionality render it a valuable tool in developing autonomous systems, particularly in the realm of multi-agent interactions and strategy adaptation.

DL, a subfield of ML, leverages complex, multi-layered artificial neural networks to model and understand diverse data structures. It has been particularly effective in tasks that involve large datasets and high-dimensional spaces. Crucial to our work is its integration in video understanding on Unmanned Aerial Vehicles (UAVs), and more particularly in Socratic Video Understanding, where the goal is to develop machines that understand video content at a human-like level. DL allows machines to learn from a series of past experiences, much like human cognition, thereby enabling them to autonomously analyze, interpret, and understand video feeds.

PID (Proportional Integral Derivative) Controllers constitute a crucial aspect of the control systems in autonomous vehicles, particularly in vehicle platooning. They calculate an ‘error’ value as the difference between a measured process variable and a desired setpoint, and apply a correction based on proportional, integral, and derivative terms to achieve the desired behavior. When combined with ML techniques and LLMs, they promise improved platooning performance, decreased fuel consumption, and better safety measures.

In the context of autonomous vehicular systems, Truck Platooning emerges as a critical application. Platooning refers to a group of vehicles that travel in close proximity to one another, operating as a single unit while improving traffic flow, reducing fuel consumption, and enhancing safety. The development of efficient, AI-enhanced control systems for platooning, therefore, holds significant implications for autonomous vehicular technology.

MASs consist of multiple interacting intelligent agents that can be used to solve problems that are difficult or impossible for an individual agent to solve. They find application across a diverse range of domains from game theory, distributed computing to robotics. In this thesis, we focus on the incorporation of LLMs in MASs, striving to achieve emergent cooperation and strategy adaptation, thereby enhancing their overall functionality and efficiency.

In a similar vein, Coevolutionary Algorithms, a type of Evolutionary Algorithms, work on the principle of co-adaptation, wherein multiple populations evolve together by interacting and competing with each other. They are particularly useful for optimization in dynamic environments, thereby finding applications in our work on strategy adaptation in MASs.

TabNet [28], a high-performance and interpretable DL model introduced by Google Cloud AI, has found growing applications in the field of tabular data learning. Its integration with Transformers, which are DL models used in understanding sequences, enhances its capacity to capture complex patterns in data. This methodology plays a critical role in our work on area estimation of forest fires.

EfficientNet [29], another state-of-the-art DL model, represents a family of advanced convolutional networks (CNNs) that leverage a compound scaling method to uniformly scale all dimensions of depth, width, and resolution. Particularly relevant in our work is its application in the early detection of glaucoma, underlining the significant potential of DL in the domain of medical imaging and diagnostics.

By amalgamating these diverse technologies, methodologies, and algorithms, this thesis delves into the expansive realm of autonomous systems. The exploration sheds light on the potential of AI-enhanced methods to address complex challenges, thus driving the advancement of autonomous systems. The synergy of these elements underscores the transformative role of AI methodologies in propelling efficiency, reliability, and security within these systems. The profound implications of these studies extend beyond their immediate applications, contributing a holistic perspective to the rapidly evolving discourse on AI and autonomous systems.

2.1 Large Language Models and Their Applications

LLMs [25, 26] such as GPT-4 by OpenAI have significantly revolutionized our approach to natural language understanding and generation. These models, by design, leverage billions of parameters and extensive corpora of text data to generate human-like text. At the heart of LLMs lie the principles of transformer architecture, characterized by self-attention mechanisms and positional encoding.

2.1.1 LLM Architecture and Mathematics

The primary mathematical model behind LLMs is the transformer model, a sequence transduction model that relies on self-attention and position-wise fully connected feed-forward networks.

A transformer model works with an encoder-decoder structure. The encoder maps an input sequence of symbol representations $(x_1, ..., x_n)$ to a sequence of continuous representations $Z = (z_1, ..., z_n)$. Given Z, the decoder then generates an output sequence $(y_1, ..., y_m)$ of symbols one element at a time, where m can be different from n.

In mathematical terms, the transformer's self-attention mechanism can be defined as:

$$\text{Self-Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.1}$$

where Q, K, V are the query, key, and value vectors. The output of the self-attention function is the weighted sum of the values, where the weight assigned to each value is computed by the compatibility function of the query with the corresponding key. Here, d_k is the dimensionality of the query and key vectors, usually set to 64 in practice.

2.1.2 LLMs Training

LLMs training is generally performed using the maximum likelihood estimation (MLE) principle. For a dataset D of N sentence pairs, where each pair (x, y) consists of an input sentence $x = (x_1, ..., x_n)$ and output sentence $y = (y_1, ..., y_m)$, the goal is to minimize the negative log-likelihood of the model parameters θ:

$$L(\theta) = -\sum_{(x,y)\in D} \log P(y|x; \theta). \tag{2.2}$$

The parameters θ of the LLM are then updated via gradient descent, specifically using optimization algorithms like Adam.

2.1.3 Challenges and Future Directions

While LLMs provide substantial advancements in natural language understanding and generation, there remain challenges such as ensuring ethical use, mitigating biases in data, and dealing with the interpretability of these models.

Despite these challenges, the potential of LLMs in various sectors is immense. With continued advancements in computational power, algorithmic improvements, and more diverse and extensive data, we can expect to see increasingly nuanced and powerful applications of LLMs in autonomous systems, healthcare, education, and more.

The unique ability of LLMs to analyze and generate human-like text, coupled with their capacity to learn from large datasets, makes them an invaluable asset in the realm of AI-enhanced autonomous systems. By continuing to innovate and explore these technologies, we can unlock unprecedented potential in machine understanding and interaction, thereby pushing the boundaries of what we perceive as possible in AI development.

2.2 Visual Language Models and Their Applications

The combination of vision and language has led to a new class of models known as Visual Language Models (VLMs) [7]. These models are trained to understand and generate information by connecting visual content with natural language, thereby allowing for more holistic and enriched interpretations of data. They have seen significant advancements with the advent of transformer architectures that effectively incorporate both vision and language.

2.2.1 VLM Architecture and Mathematics

A VLM consists of two main components: a vision model and a language model. These components are typically connected through a transformer layer that maps the vision and language representations into a common semantic space.

The vision model is usually a Convolutional Neural Network (CNN) that extracts feature vectors from input images. For a given image I, a CNN will output a feature map $F = \text{CNN}(I)$, where $F \in \mathbb{R}^{h \times w \times d}$, h and w are the height and width of the feature map, and d is the dimensionality of the feature vectors.

The language model is a transformer-based language model, which generates a sequence of token embeddings for a given text input. For an input sentence S consisting of n tokens, the language model outputs a sequence of token embeddings $E = \mathrm{LM}(S)$, where $E \in \mathbb{R}^{n \times d'}$ and d' is the dimensionality of the token embeddings.

These two components are combined using a transformer layer, which connects the vision and language representations:

$$O = \mathrm{Transformer}(F, E), \tag{2.3}$$

where $O \in \mathbb{R}^{(h \times w + n) \times d''}$, and d'' is the dimensionality of the output representations.

2.2.2 VLMs Training

Training a VLM involves learning a joint distribution over images and text. In practice, this is usually done through contrastive learning. The objective is to maximize the similarity between the representations of a pair of positive examples (an image and its corresponding text), and minimize the similarity between the representations of negative examples (an image and random text, or vice versa).

Given a batch of B image-text pairs $(I_z, S_z)_{z=1}^{B}$, the contrastive loss can be defined as:

$$L = -\frac{1}{B} \sum_{z=1}^{B} \log \frac{\exp(s_{zz}/\tau)}{\exp(s_{zz}/\tau) + \sum_{o \neq z} \exp(s_{zo}/\tau)}, \tag{2.4}$$

where $s_{zo} = \mathrm{sim}(I_z, S_o)$ is the similarity between image I_z and text S_o, and τ is a temperature parameter.

2.2.3 Challenges and Future Directions

Although VLMs have demonstrated promising results in various tasks, there are still several challenges to be addressed, such as mitigating biases in data, improving the interpretability of these models, and enhancing their efficiency.

Moreover, a new wave of VLMs is emerging, with models like BLIP-2, which aim to further improve the performance and utility of VLMs in real-world applications. These next-generation models, incorporating more complex and powerful architectures, are paving the way for unprecedented advancements in AI.

VLMs, with their ability to merge vision and language understanding, play a critical role in the development of intelligent and interactive AI systems. They provide a powerful tool for enhancing autonomous systems, contributing significantly to the field of AI research and development.

2.3 PID Control and Its Applications

PID Control is one of the most commonly used feedback control mechanisms in various fields due to its simple, effective and comprehensible control structure. It has been employed to regulate everything from industrial processes to vehicle control, and even in some facets of AI applications.

2.3.1 The Mathematics of PID Control

The PID controller is named after its three correcting terms, each of which respectively corrects the present, accumulated past, and future behavior of the error.

The control function of a PID controller can be defined as follows:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt}, \tag{2.5}$$

where:

- $u(t)$ is the output control variable.
- K_p, K_i, and K_d are the proportional, integral, and derivative gains respectively.
- $e(t)$ is the error signal defined as $e(t) = r(t) - y(t)$, with $r(t)$ being the desired setpoint and $y(t)$ being the actual output.

The PID control law is then composed of three terms:

1. The proportional term (P) is proportional to the current error $e(t)$. It determines the reaction to the current error. The proportional response can be adjusted by K_p.
2. The integral term (I) is proportional to both the magnitude and the duration of the error. It determines the reaction based on the sum of recent errors. The integral response can be adjusted by altering K_i.
3. The derivative term (D) is proportional to the rate of change of the error. This predicts the future trend of the error, based on its current rate of change. The impact of the derivative response can be adjusted by changing K_d.

2.3.2 Tuning of PID Controllers

Tuning a PID controller involves adjusting the proportional, integral, and derivative gains to get the desired response. There are several methods for PID controller tuning, with the most popular ones being the methods Ziegler–Nichols and Cohen–Coon [17].

The goal is to find the gain values that minimize the difference between the desired and actual output. This is often formulated as an optimization problem, which can be solved using various techniques.

2.3.3 Applications and Challenges

Despite its apparent simplicity, PID control is a cornerstone of control engineering, applied widely in various fields [18]. However, tuning a PID controller for a specific system can be challenging, especially for systems with complex and nonlinear dynamics [19]. Furthermore, while PID controllers are effective for systems with constant parameters, they may fail to deliver satisfactory control performance for systems with varying parameters or disturbances.

Incorporating AI techniques like ML and optimization algorithms can significantly improve the performance of PID controllers. With the advent of AI-enhanced PID controllers, we can now handle more complex control tasks and navigate dynamic, uncertain environments more effectively.

2.4 Truck Platooning

Truck Platooning involves the application of automated driving technology to enable a group of trucks to travel in a close convoy [16]. This method increases fuel efficiency and safety, but also necessitates intricate control systems for managing inter-vehicle distances and reacting to real-world variables. The integration of LLMs into this context can lead to improved explainability and control.

A conventional approach to truck platooning involves the use of control algorithms such as the PID control. This controller manipulates the system's output, i.e., the acceleration of the truck, based on the error in the system's desired output, which is the desired inter-vehicle distance. If d_{ref} is the desired inter-vehicle distance and d is the current distance, then the error e is given by

$$e = d_{\text{ref}} - d. \tag{2.6}$$

The output of the PID controller u is given by

$$u = K_p e + K_i \int e dt + K_d \frac{de}{dt}, \tag{2.7}$$

where K_p, K_i, and K_d are the proportional, integral, and derivative gains, respectively.

In a traditional setup, these gains are usually set by the engineer during system design and are fixed. However, we propose to use a DL model to predict the gains based on real-time data, resulting in an adaptive PID controller.

Assuming a simplified linear model of the truck dynamics, the state of the truck can be represented by its position p and velocity v. If we assume that $O = \{DL, LLM, \ldots\}$, where this can be any generic AI model, takes as input the state of the truck and the error e, and outputs the gains K_p, K_i, and K_d, then we can represent this as

$$[K_p, K_i, K_d] = O(p, v, e). \tag{2.8}$$

This results in an adaptive PID controller that can adjust its behavior based on real-time conditions.

2.4.1 Explainability with LLMs

In addition to control, LLMs can also enhance the explainability of the platooning system. By analyzing the inputs and outputs of the LLM, we can gain insights into the decisions made by the controller. For example, the LLM might place more emphasis on the proportional term when the error is large, indicating that it is focusing on reducing the error quickly. On the other hand, when the error is small and changing slowly, it might place more emphasis on the integral term, indicating that it is trying to eliminate the steady-state error. This can help in understanding the behavior of the controller and making adjustments if necessary.

In conclusion, the incorporation of LLMs into truck platooning can enhance not just the control performance, but also the explainability of the system. However, practical implementation of this approach requires thorough testing and validation to ensure safety and reliability.

2.5 Multi-agent Systems and Coevolutionary Theory

A MAS is a computational system where several autonomous entities, called agents, interact or work together to perform tasks or solve problems that might be too complex or large for a unique agent. These agents can exhibit complex behaviors and evolve through their interactions in the system.

The mathematical model of a MAS can be formalized as a tuple $MAS = \{A, Env, \rho, Ag\}$, where:

- $A = \{a_1, a_2, ..., a_n\}$ is a set of agents.
- Env represents the environment in which the agents interact.
- $\rho : A \times Env \to A$ is a function describing the rules of agent behavior in the environment.
- $Ag : A \times A \to A$ represents the aggregation function describing how the agents interact with each other.

2.5.1 Introduction to Coevolutionary Theory

Coevolutionary theory [42] is a key concept in the study of evolutionary processes. It pertains to the change of a biological entity triggered by the change of a related entity. In the context of MAS, coevolutionary theory implies the concurrent evolution of agents based on their interactions and the continuous adaptation to each other's strategy.

2.5.2 Coevolution in Multi-agent Systems

In the context of a MAS, a coevolutionary algorithm could involve a population of solutions for each agent, evolving concurrently. The fitness of a solution is then determined not just by its own characteristics, but also by the characteristics of the other agents' solutions in the environment.

Let's denote a solution for agent a_z as s_{a_z}, and the set of all solutions for this agent as $S_{a_z} = \{s_{a_z1}, s_{a_z2}, ..., s_{a_zm}\}$. The fitness $F_{a_z}(s_{a_z})$ of a solution s_{a_z} can be defined by its interaction with the solutions of the other agents:

$$F_{a_z}(s_{a_z}) = \sum_{o=1, o\neq z}^{n} \int_{s_{a_o} \in S_{a_o}} f_{zo}(s_{a_z}, s_{a_o}) ds_{a_o}, \tag{2.9}$$

where $f_{zo}(s_{a_z}, s_{a_o})$ is a function describing the outcome of the interaction between solution s_{a_z} of agent a_z, and solution s_{a_o} of agent a_o.

2.5.3 Challenges and Opportunities

Implementing coevolution in MAS involves several challenges, including maintaining diversity, avoiding premature convergence, and dealing with the problem of relativism, i.e., defining an absolute measure of performance in a coevolutionary setup. Despite these challenges, coevolution offers a powerful framework to harness the emergent behavior and adaptation capabilities of MAS, making it a promising approach for designing efficient and robust distributed systems.

2.5.4 LLMs in Coevolutionary Multi-agent Systems

As we have seen in the context of individual agents, LLMs can imbue agents with enhanced capabilities, opening up interesting possibilities for coevolution in MAS. In such a scenario, agents would evolve not just based on predefined rule sets, but also through the strategies inferred from the LLMs.

One approach to integrating LLMs into coevolutionary MAS involves the use of these models as strategy generators. In this case, the LLM would generate possible strategies or actions, and these would be evaluated within the coevolutionary context. For agent a_z, this could be formalized by introducing a strategy generation function G_{a_z} that maps the agent's current state to a strategy space, Ω_{a_z}:

$$G_{a_z} : S_{a_z} \to \Omega_{a_z}. \tag{2.10}$$

The fitness function can then be redefined to account for these generated strategies:

$$F_{a_z}(s_{a_z}) = \sum_{o=1, o \neq z}^{n} \int_{s_{a_z} \in S_{a_z}} f_{zo}(G_{a_z}(s_{a_z}), G_{a_o}(s_{a_o})) ds_{a_o} \tag{2.11}$$

Here, the interaction function f_{zo} is evaluated based on the generated strategies $G_{a_z}(s_{a_z})$ and $G_{a_o}(s_{a_o})$.

This use of LLMs can allow for richer dynamics in the coevolutionary process, as strategies can be more diverse and adaptive, potentially leading to more sophisticated emergent behaviors. Furthermore, the use of an LLM allows for the encoding of complex, human-like strategies, which can add an additional layer of complexity and potential performance in the MAS. This novel approach to coevolutionary MAS could potentially lead to the discovery of new, more efficient strategies and solutions. However, further research is needed to determine the specific mechanics of such integration and its benefits.

2.6 Socratic Video Understanding

Socratic Video Understanding (SVU) [22] is a methodology for interpreting video content and acting upon, where several modules exchange information between them. It involves using AI models to analyze a video and make predictions or draw conclusions about the content. In this work, we explore the use of both LLMs and VLMs to perform this task, as their size and complexity allow for a deeper understanding of the content and its context.

LLMs, such as GPT-3, have been used to perform a variety of language understanding tasks, including the comprehension of a sequence of events, context, and other complex language-based tasks. By applying these models to video understanding, we can gain insights into the events, characters, and objects within a video.

Suppose we have a video V, and we represent the sequence of frames in the video as $F_1, F_2, ..., F_n$. We can use an image-to-text model M_{img2txt} to generate a textual description D_z of each frame F_z, i.e., $D_z = VLM_{\text{img2txt}}(F_z)$. Then, the LLM can be used to understand the sequence of descriptions, which is equivalent to understanding the video.

$$LLM(D_1, D_2, ..., D_n). \tag{2.12}$$

Through this methodology, we can obtain a nuanced understanding of the video content.

2.7 TabNet

TabNet [28] is a novel DL model developed for tabular data. The model, unlike traditional DL architectures, can handle heterogeneous data and enables interpretability through learned feature importance. Its architecture can be represented mathematically as follows:

$$f_{\text{TabNet}}(x) = \sum_{k=1}^{K} M_k \cdot T_k(x), \tag{2.13}$$

where x is the input data, K is the number of decision steps, M_k are the learnable masks, and $T_k(x)$ are the transformers encoding the input data x at step k.

2.7.1 Application to Forest Fire Area Estimation

The determination of affected areas during forest fires is a crucial task for resource allocation and the planning of effective firefighting strategies. In this thesis, we employ TabNet to tackle this challenge, proving its effectiveness in accurately predicting the extent of forest fires.

In forest fire area estimation, the input data x typically includes meteorological data (temperature, humidity, wind speed, etc.), topographical data (elevation, slope, etc.), and other relevant information such as the type of vegetation.

For the k-th decision step, TabNet applies the learned mask M_k to the transformed input $T_k(x)$ to select relevant features. By allowing TabNet to learn the importance of different features at each step, the model is capable of identifying the most relevant factors that contribute to the spread of a forest fire.

To train the model, we define the objective function as the Mean Squared Error (MSE) between the predicted and true fire areas:

$$L(\Theta) = \frac{1}{n} \sum_{z=1}^{n} \left(f_{\text{TabNet}}(x_z; \Theta) - y_z \right)^2 , \tag{2.14}$$

where Θ represents the parameters of the TabNet model, x_z is the z-th input data, and y_z is the true fire area corresponding to x_z. The objective of training is to minimize this loss function:

$$\Theta^* = \arg\min_{\Theta} L(\Theta). \tag{2.15}$$

Through this approach, we leverage the strengths of TabNet, particularly its ability to handle heterogeneous data and provide interpretability, to achieve accurate and insightful forest fire area estimation.

2.8 EfficientNet

EfficientNet [29] is a DL model for image classification that optimizes the balance between the depth, width, and resolution of the network. It is based on a compound scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet effective compound coefficient. Its architecture can be represented as follows:

Let d, w, and r represent the depth, width, and resolution of a baseline network, respectively. The scaled network depth D, width W, and resolution R are then computed as follows:

$$D = d \cdot \alpha^{\phi}, \quad W = w \cdot \beta^{\phi}, \quad R = r \cdot \gamma^{\phi}, \tag{2.16}$$

where ϕ is a user-specified coefficient that controls the resources available for model scaling, and α, β, and γ are constants that can be determined by a small grid search.

2.8.1 Application to Glaucoma Detection

Glaucoma is a complex eye disease that can lead to irreversible blindness if not detected and treated early. In this thesis, we employ EfficientNet to address this crucial task.

In glaucoma detection, the input data typically includes retinal images obtained from ocular examinations. These images are preprocessed and resized to the specified resolution R before being fed into the model EfficientNet.

We use the architecture of the EfficientNet to automatically learn features from the input retinal images. These features can include various aspects of the optic nerve head, such as the cup-to-disc ratio and the thickness of the retinal nerve fiber layer, which are critical for diagnosing glaucoma.

The model is trained to minimize the cross-entropy loss between the predicted and true labels:

$$L(\Theta) = -\frac{1}{n} \sum_{z=1}^{n} \left[y_z \log(p_\Theta(y_z|x_z)) + (1 - y_z) \log(1 - p_\Theta(y_z|x_z)) \right], \tag{2.17}$$

where Θ represents the parameters of the model EfficientNet, x_z is the z-th input retinal image, y_z is the true label corresponding to x_z (1 for glaucoma and 0 for normal), and $p_\Theta(y_z|x_z)$ is the predicted probability of the z-th image being classified as glaucoma. The objective of training is to find the parameters Θ that minimize this loss:

$$\Theta^* = \arg\min_{\Theta} L(\Theta). \tag{2.18}$$

Through this approach, we leverage the strengths of EfficientNet, particularly its efficiency and performance, to achieve accurate and early detection of glaucoma.

2.9 Technological Framework and Application Justification

In this thesis, we establish a distinct methodological approach where each technology – LLMs, DL, and EC Theory – is applied within specific application fields. The rationale behind these applications and their corresponding AI approaches are detailed below.

2.9.1 Application of LLMs in Autonomous Systems

LLMs like GPT-3.5-turbo are deployed in autonomous vehicle platoons to enhance real-time decision-making and communication. The choice of LLMs for this application is predicated on their ability to process and generate natural language, enabling vehicles to understand and respond to complex commands and to communicate with one another in an interpretable manner.

- **Objective:** To improve coordination and response strategies within platooning systems as well as explainability, thus enhancing overall safety and efficiency.
- **Justification:** The robust natural language processing capabilities of LLMs provide a means for nuanced communication and sophisticated control strategies beyond the scope of traditional algorithms.

2.9.2 DL for Scene Understanding and Medical Diagnostics

DL techniques are harnessed for semantic scene understanding in UAVs and for medical diagnostics in the early detection of glaucoma.

- **Objective:** To interpret complex visual data streams and to identify pathological markers in medical imagery, respectively.
- **Justification:** The convolutional neural networks (CNNs) at the heart of DL are exceptionally suited for image recognition tasks, capable of identifying patterns and features that are imperceptible to human analysts.

2.9.3 EC Theory in Strategy Formation

EC is applied within MASs to establish a game-theoretic framework where we simulate and analyze the dynamics of cooperation and defection.

- **Objective:** To model emergent behavior and to optimize collective strategies in dynamic environments.
- **Justification:** The adaptive nature of the EC Theory allows for the evolution of agent behaviors that can dynamically adjust to complex and changing conditions, reflecting the unpredictable nature of real-world interactions.

2.9.4 Cross-disciplinary Applications

The confluence of these AI technologies is justified by the complementary strengths they bring to their respective fields:

- In the domain of **autonomous vehicles**, LLMs and DL are combined to enable vehicles to interpret sensor data, communicate with one another, and make decisions in a human-like manner.
- For **UAV scene comprehension**, DL provides the visual processing capabilities, while LLMs contribute to the understanding and generation of descriptive narratives about the scenes.
- In **strategy formation within MASs**, EC sets a game-theoretic framework of utility maximization, while LLMs enrich the strategic options with their ability to generate a diverse array of potential actions.

Each application domain is carefully selected based on the inherent demands of the field and the suitability of the AI technology to meet those demands. The integration of AI approaches is not arbitrary but is instead a deliberate choice to harness the strengths of each technology, thereby creating systems that are more than the sum of their parts.

The methodology of this thesis is anchored in the belief that the future of autonomous systems lies in the synergy of multiple AI technologies. By delineating the specific roles and applications of LLMs, DL, and the EC Theory, we set the stage for a comprehensive exploration of their potential and pave the way for their sophisticated integration in complex real-world scenarios.

Chapter 3

LLM Adaptive PID Control for B5G Truck Platooning Systems

I. de Zarzà, J. de Curtò, Gemma Roig and Carlos T. Calafate. (2023). "LLM Adaptive PID Control for B5G Truck Platooning Systems" Sensors, vol(23), 5899.
DOI: 10.3390/s23135899

This chapter presents an exploration into the capabilities of an adaptive PID controller within the realm of truck platooning operations, situating the inquiry within the context of Cognitive Radio and AI-enhanced 5G and Beyond 5G (B5G) networks. We developed a Deep Learning (DL) model that emulates an adaptive PID controller, taking into account the implications of factors such as communication latency, packet loss, and communication range, alongside considerations of reliability, robustness, and security. Furthermore, we harnessed a Large Language Model (LLM), GPT-3.5-turbo, to deliver instantaneous performance updates to the PID system, thereby elucidating its potential for incorporation into AI-enabled radio and networks. This research unveils crucial insights for augmenting the performance and safety parameters of vehicle platooning systems within B5G networks, concurrently underlining the prospective applications of LLMs within such technologically advanced communication environments.

3.1 Introduction and Related Work

The concept of truck platooning [1, 2, 3, 4] is gaining significant attention due to its potential to improve fuel efficiency, reduce traffic congestion, and enhance road safety [5, 6]. In a platoon, multiple trucks travel closely together, maintaining a constant distance to minimize air drag and save fuel. Adaptive PID (Proportional-Integral-Derivative) controllers play a crucial role in maintaining a constant inter-vehicle distance and ensuring the stability of the platoon. By adjusting the controller gains in real time based on the system's behavior, adaptive PID controllers [7] can enhance the performance of the platoon and adapt to various driving conditions.

Effective communication between vehicles is fundamental for the successful implementation of truck platooning. Vehicle-to-vehicle (V2V) [8, 9] communication enables trucks to share vital information, such as the speed, position, and braking status, with other vehicles in the platoon. This information is essential for maintaining a safe and constant distance between the trucks, which ensures an efficient platoon operation. The key aspects of communication that impact the performance of a platoon include the communication latency, packet loss, communication range, reliability, and robustness. Moreover, the security of the communication system is of the utmost importance, as it protects the platoon from potential cyberattacks and ensures the safety of the drivers and the cargo.

In recent years, the development of Cognitive Radio and AI-enabled 5G and Beyond 5G (B5G) networks has opened up new opportunities for advanced vehicular communication systems [12, 10, 11]. Indeed, one such application is truck platooning [14, 13, 15]. Our goal with this work is to emphasize the role of AI-enabled radio and networks in enhancing communication between vehicles, thereby addressing key challenges.

This chapter presents an adaptive PID controller [16, 17, 18] that utilizes a Deep Learning (DL) model for efficient and reliable truck platooning. The controller is designed to handle various aspects of vehicle-to-vehicle (V2V) communication, such as communication latency, packet loss, communication range, reliability, and robustness. Furthermore, security concerns [20, 19] are addressed to ensure the safety of the platoon.

We begin with a base adaptive PID controller [21, 22] that leverages a trained neural network model to predict the actual inter-vehicle distance. The controller is then improved by incorporating considerations relating to communication latency, packet loss, and communication range, as we believe in the importance of a reliable and robust communication system for the safe operation of the platoon. Moreover, a cutting-edge Large Language Model (LLM) [23, 24, 25] , GPT-3.5-turbo, is integrated into the system to obtain real-time performance updates, demonstrating an innovative application of LLMs in the context of truck platooning. The results indicate that our adaptive PID controller, along with the LLM-based performance updates, offers a promising solution for efficient and secure truck platooning. An illustration of the proposed system is depicted in Figure 3.1.

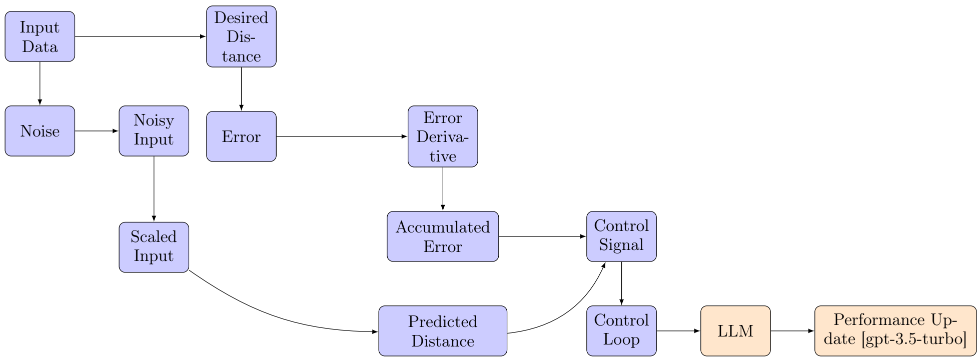


Figure 3.1: Flow diagram of the adaptive PID controller with LLM performance updates.

This research aims to provide valuable insights into the design and implementation of AI-driven control systems for truck platooning in B5G networks while highlighting the potential of LLMs in advanced communication environments.

The core contributions of this study are twofold. Firstly, we put forth an adaptive PID controller, aimed at bolstering the efficiency and reliability in truck platooning within AI-enabled 5G and B5G network contexts. This controller deploys a DL model to forecast actual inter-vehicle distances and introduces factors such as communication latency, packet loss, communication range, and system reliability and robustness. These enhancements are designed to augment the performance and safety of the platoon amid varied driving conditions. Secondly, we introduce the integration of a cutting-edge Language Learning Model (LLM), GPT-3.5-turbo, into the control loop. The LLM provides real-time updates and recommendations, thereby augmenting the adaptability and explainability of the PID controller. This innovative usage of LLMs within the realm of truck platooning allows the system to tap into natural language comprehension abilities, which then leads to improved decision making and system optimization. Taken together, the adaptive PID controller and the LLM integration represent a comprehensive solution that guarantees the effective and secure truck platooning in B5G networks.

The remainder of this chapter is organized as follows: Section 3.2 introduces the adaptive PID controller, emphasizing its critical role in control systems, with particular attention given to vehicle platooning. Section 4.4 outlines the methodology employed to design and implement the adaptive PID controller, addressing key considerations and challenges such as the latency, packet loss, communication range, noise channel, and security. Subsequently, Section 4.3.4 explores the integration of LLMs as a means to enhance the performance of the adaptive PID controller through real-time updates and recommendations. In Section 4.7, we delve into the potential implications of our research. Finally, Section 4.8 concludes the chapter and outlines potential directions for future research.

3.2 Adaptive PID Controller

A Proportional-Integral-Derivative (PID) controller is a widely used control algorithm in various control systems. It calculates the control signal based on the error, the integral of the error, and its derivative. The error ($e(t)$) is the difference between the desired setpoint ($r(t)$) and the measured process variable ($y(t)$):

$$e(t) = r(t) - y(t). \tag{3.1}$$

The control signal $u(t)$ generated by the PID controller is given by the following equation:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt}, \tag{3.2}$$

where K_p, K_i, and K_d are the proportional, integral, and derivative gains, respectively.

In an adaptive PID controller, the gains K_p, K_i, and K_d are adjusted in real time based on the system's performance. The goal is to maintain optimal control performance despite changes in the system dynamics or external disturbances. Various methods exist for tuning the PID gains adaptively, such as Ziegler–Nichols [26], Cohen–Coon, and model-based approaches [16]. In this work, we employed a data-driven approach by training a deep neural network (DNN) to predict the optimal PID gains for the given system state. The DNN was trained on a synthetic dataset that captured a wide range of system behaviors and conditions and was used to illustrate its practical application. This allowed the adaptive PID controller to adjust the gains in real time based on the system's current state, thus ensuring optimal control performance.

This chapter primarily centers on the creation and assessment of an adaptive DNN-based PID control methodology for truck platooning. Although the synthetic dataset we utilized for experimentation did not overtly include vehicle model parameters such as mass, inertia, aerodynamic drag, tire friction, and powertrain features, it is crucial to stress that our approach can still be implemented when these parameters are either known or can be estimated. The synthetic dataset was representative, and its use demonstrated the efficacy and performance of our adaptive control method.

The adaptive PID controller employs the strength of deep neural networks to understand and adapt to the inherent dynamics of the truck platooning system. This enables it to effectively manage the changes in vehicle characteristics and driving conditions. This work propels the progression of adaptive control strategies within the scope of truck platooning, setting the stage for future research that could merge intricate vehicle models with parameter estimation methodologies.

The DNN used in this work comprised three types of layers: the input, hidden, and output layers. The input layer accepted the normalized system state, and the output layer generated the predicted PID gains. The hidden layers contained multiple neurons with activation functions, which facilitated the learning of complex nonlinear relationships between the input and output. Specifically, our DNN architecture included two hidden layers, the first with 64 neurons and the second with 32 neurons, both using the Rectified Linear Unit (ReLU) activation function. To prevent overfitting, we also incorporated dropout layers with a dropout rate of 0.2.

We adopted a supervised learning approach to train the DNN and utilized a synthetic dataset comprising various system states and their corresponding optimal PID gains. The dataset was generated using different parameters, such as desired and actual distances, vehicle speed, acceleration, road grade, and weather conditions. The dataset was then divided into training and testing sets, using an 80/20 split.

To optimize the DNN, we minimized the mean square error (MSE) loss function, which measures the difference between the predicted PID gains and the true optimal gains. For this purpose, we employed the Adam optimizer with a learning rate of 0.001. The training process also included a validation split of 20%, with the model being trained for 50 epochs using a batch size of 32; the training and validation loss over time is shown in Figure 3.2. Once the DNN is trained, it can predict optimal PID gains for new unseen system states, enabling the adaptive PID controller to adjust its gains in real time and maintain optimal control performance.

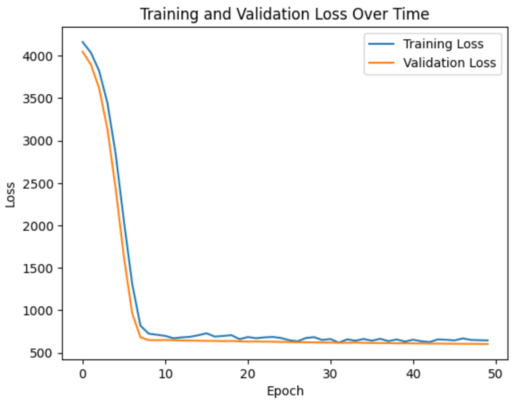


Figure 3.2: Training and validation loss over time for the proposed architecture.

The trained DNN, shown in Figure 3.3, was integrated into the control loop of the truck platooning system. At each time step, the current system state was passed as input to the DNN, which predicted the optimal PID gains. These gains were then used to calculate the control signal, which adjusted the truck's acceleration or deceleration to maintain a safe and constant inter-vehicle distance. This adaptive approach allows the PID controller to respond effectively to changes in the system dynamics and external disturbances, ensuring the stable and efficient operation of the platoon.

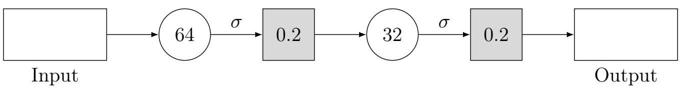


Figure 3.3: Detailed network architecture of the deep neural network (DNN) used for the adaptive PID tuning. The model consists of two fully connected layers with 64 and 32 neurons, followed by dropout layers with a rate of 0.2. Rectified Linear Unit (ReLU) activation functions, depicted as σ, are applied after each fully connected layer. The input and output layers are also depicted.

This study, while recognizing the importance of stability analysis, did not undertake a comprehensive formal stability examination via strict mathematical methods. We acknowledge that such an analysis is a fundamental facet of controller design, especially within the realm of adaptive control systems; the focus of our work, however, veered towards the implementation of LLMs for the sake of explainability.

Stability analysis in adaptive control systems typically involves an examination of the closed-loop system's stability attributes, a task rendered complex by the controller's adaptive properties. Traditional stability analysis methods, such as the Lyapunov stability theory and small-gain theorems, are usually employed to scrutinize the stability of adaptive control systems. These methods often require the establishment of suitable Lyapunov functions or the study of system gains to ensure stability and convergence.

The intricate nature and rigorous mathematical demand of a complete stability analysis meant that it was beyond the scope of our current work. Future research may seek to perform an in-depth stability analysis to provide formal assurances and further substantiate the stability properties of the adaptive PID controller under various scenarios. The main thrust of this chapter, however, is to explore and highlight the utility of LLMs for the enhancement of system explainability.

3.3 Methodology

For our simulation, we utilized a platoon of two trucks that maintained a safe inter-vehicle distance. The synthetic dataset under study considered a desired distance between the two trucks to be within the range of 20–100 m (although in commercial applications, the range would be much lower in order to benefit from the aerodynamic drag), with their speeds varying between 40 and 120 km/h. The safe distance between the vehicles was determined based on various factors, such as the trucks' speeds, acceleration, road grade, and weather conditions. The control loop calculated the control signal based on the current state, which was then used to update the truck's acceleration or deceleration.

The choice of varying the desired distance between 20 and 100 m was meant to simulate different traffic scenarios and communication conditions that may affect the performance of the platoon control system. In real-world applications, the desired distance between vehicles might not necessarily remain constant, as factors such as traffic density, road conditions, and safety considerations can influence the optimal distance.

To simulate the communication latency, we modified the control loop to include a circular buffer for the control signals. This buffer represents the delay in communication between vehicles, with each element in the buffer corresponding to a time step of latency.

We set the communication latency (in time steps) and initialized the control signal buffer accordingly. The desired distances, actual distances, and control signals were recorded for each time step. After running the simulation, we visualized the desired and actual inter-vehicle distances, as well as the control signals over time. The plots in Figures 4.2 and 4.3 can help in the analysis of the latency's impact on the performance of the adaptive PID controller and its ability to maintain safe inter-vehicle distances.

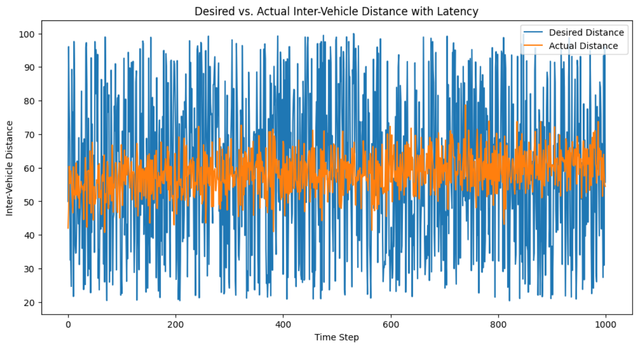


Figure 3.4: Desired vs. actual inter-vehicle distance with latency.

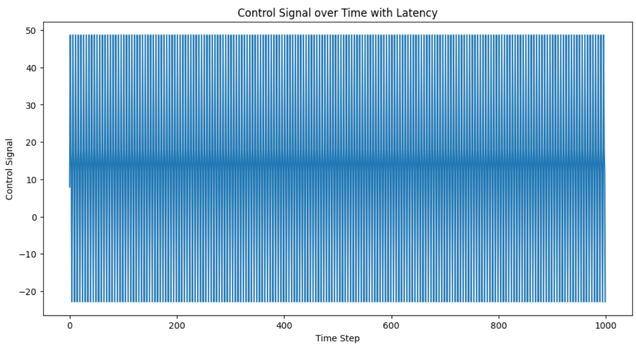


Figure 3.5: Temporal evolution of the control signal amid the latency. The control signal encapsulates the system modifications applied to sustain the requisite distance between the vehicles within a truck platoon, acting as a responsive adjustment to the latency-induced variations.

Thus, communication latency was integrated into the control loop by simulating a circular buffer for the control signals. This modification, which meant that the effects of latency were now incorporated into the control loop, allowed us to analyze the performance of the adaptive PID controller under various latency conditions. In the control loop, the buffer was used to store and retrieve control signals with the specified latency. At the beginning of each iteration, a placeholder was added to the buffer, and the delayed control signal was retrieved by popping the first element. If the delayed control signal was available, it was used to control the vehicle; otherwise, the current PID calculation was used. This process simulated communication latency and helped us to understand its impact on the system's performance. In this specific example, a latency of five time steps was used.

To simulate the packet loss in the communication between vehicles, we modified the control loop to incorporate a packet loss rate. This rate represents the percentage of control signals that are lost during transmission. The packet loss was simulated by randomly setting the percentage of control signals to None based on the packet loss rate. Mathematically, we can define the packet loss rate as $p \in [0, 1]$, where $p = 0$ means no packet loss, and $p = 1$ means 100% packet loss. We then generated a random number $r \in [0, 1]$ for each time step, and if $r < p$, we set the control signal to None. We subsequently set the packet loss rate and ran the simulation, recording the desired distances, actual distances, and the control signals for each time step. After running the simulation, we visualized the desired and actual inter-vehicle distances, as well as the control signals over time. The plots in Figures 4.4 and 3.7 can help in analyzing the impact of packet loss on the performance of the adaptive PID controller and its ability to maintain safe inter-vehicle distances.

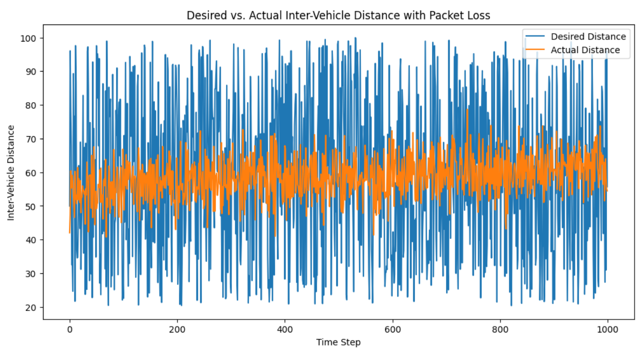


Figure 3.6: Desired vs. actual inter-vehicle distance with packet loss.

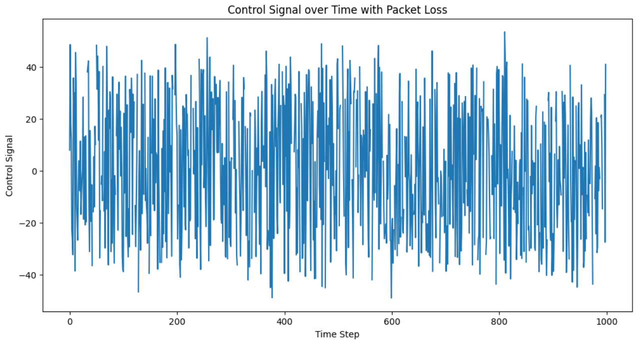


Figure 3.7: Control signal trajectory amid the packet loss. This illustrates the control signal's role as a corrective mechanism that dynamically adjusts to maintain the intended inter-vehicle distance within a truck platoon, demonstrating its resilience despite the packet loss events.

In the specific example, the packet loss rate was set to 0.1 (or 10%). The code can also be used to test a range of packet loss rates to understand the sensitivity of the controller's performance under different packet loss scenarios.

However, in a real environment, the reality is that packet loss gradually increases as the distance increases, up until it reaches 100%. The all-or-nothing approach used in the previous code might not accurately represent this behavior. To better simulate the real-world scenario, we modified the control loop to incorporate a gradual increase in packet loss as the distance increased. Our approach involved the use of a sigmoid function, illustrated in Figure 3.8, to map the distance to a packet loss rate (see Figures 3.9 and 3.10).

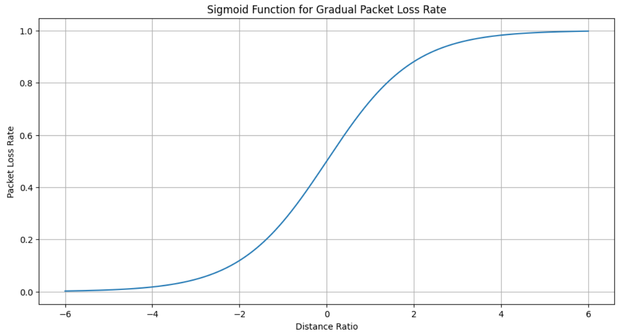


Figure 3.8: Sigmoid function for gradual packet loss rate. This plot illustrates the relationship between the distance ratio (predicted distance divided by communication range) and the packet loss rate. The sigmoid function demonstrates a gradual increase in packet loss rate as the distance ratio increases, simulating a more realistic communication scenario in the control loop.

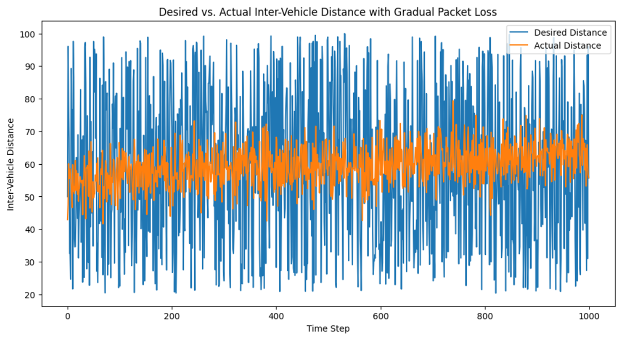


Figure 3.9: Desired vs. actual inter-vehicle distance with a gradual packet loss.

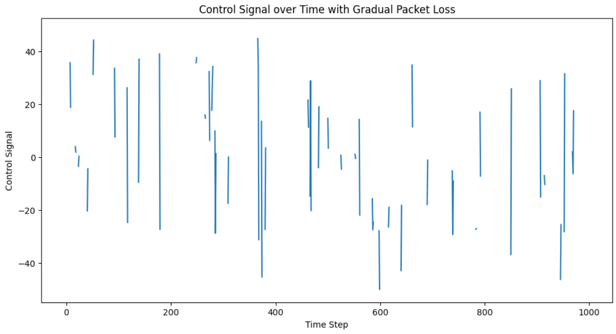


Figure 3.10: Temporal progression of the control signal amid gradual packet loss. The control signal functions as an adaptive mechanism that continually adjusts to preserve the targeted distance between vehicles within a truck platoon, even when confronting the challenges of a gradual packet loss.

In our study, we introduced the gradual packet loss mechanism based on the sigmoid function to simulate a more realistic scenario where packet loss increases as the inter-vehicle distance approaches the maximum communication range. Although the sigmoid function indeed resulted in a 100% packet loss rate when the maximum communication range was reached, it was essential to conduct experiments in order to investigate the system's behavior under varying communication conditions and packet loss rates.

The purpose of these experiments was to demonstrate the performance and robustness of the proposed control strategy in maintaining the desired inter-vehicle distance despite the presence of communication challenges. Figure 3.9, which presents the distances between trucks, may not show significant differences as compared to the previous results. However, it is crucial to highlight the controller's capability to handle communication limitations and maintain satisfactory performance even when the vehicles were close to or at the communication range's limits. This observation underscores the importance of analyzing the impact of communication range and packet loss on the control system's performance in real-world applications.

In order to simulate the effect of communication range limitations on the adaptive PID controller, we modified the control loop to take into account the communication range. When the predicted inter-vehicle distance was greater than the communication range, the control signal was set to None, which simulated a lack of communication between the vehicles. We then set the communication range and ran the simulation, recording the desired distances, actual distances, and control signals for each time step. After running the simulation, we visualized the desired and actual inter-vehicle distances, as well as the control signals over time. The plots in Figures 3.11 and 3.12 help in analyzing the impact of communication range limitations on the performance of the adaptive PID controller and its ability to maintain safe inter-vehicle distances.

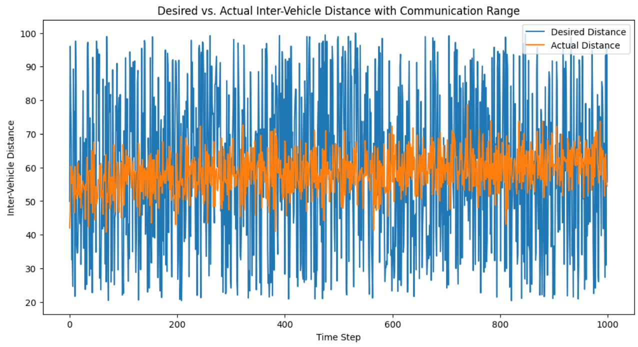


Figure 3.11: Desired vs. actual inter-vehicle distance with the communication range.

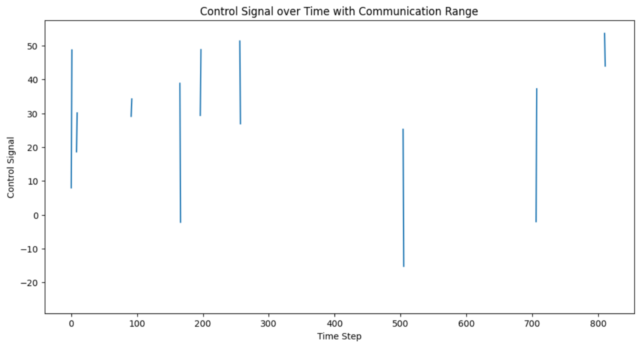


Figure 3.12: Control signal trajectory in varying communication ranges. The control signal, depicted here, acts as a real-time corrective measure that effectively regulates inter-vehicle distance within a truck platoon, demonstrating its adaptability across different communication range scenarios.

To evaluate the effect of noisy communication on the adaptive PID controller, we modeled the impact of Gaussian noise on the packet loss, which influences the system's ability to accurately calculate the control signal. Let $\mathcal{N}(0, \sigma^2)$ be the Gaussian noise, with mean 0 and standard deviation σ. We incorporated the noise effect by mapping the noise standard deviation to a packet loss rate using a sigmoid function. We then ran the control loop with the noise-affected packet loss rate and recorded the desired distances, actual distances, and control signals for each time step. After running the simulation, we visualized the desired and actual inter-vehicle distances, as well as the control signals over time. The plots in Figures 3.13 and 3.14 help in analyzing the performance of the adaptive PID controller under noisy communication conditions and its ability to maintain safe inter-vehicle distances.

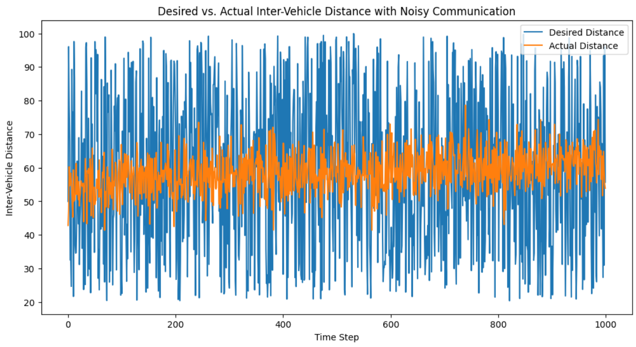


Figure 3.13: Desired vs. actual inter-vehicle distance with noisy communication.

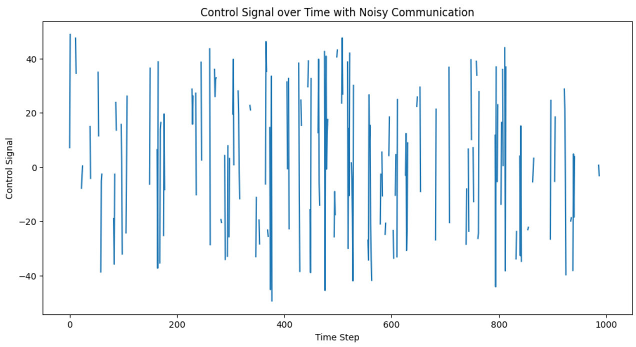


Figure 3.14: Control signal behavior amid noisy communication. This depiction of the control signal underscores its role as a dynamic corrective measure, adjusting in real time to manage inter-vehicle distances within a truck platoon, even under the challenging conditions of communication noise.

The reason behind adding Gaussian noise to the packet loss rate is the following: While it is true that the sigmoid function models the effect of signal attenuation on packet loss, we wanted to explore the effect of other sources of noise that could also impact the communication range, such as atmospheric conditions or interference from other wireless signals. By adding Gaussian noise to the packet loss rate, we introduced a random component to the simulation that could help us better understand the robustness of the platooning system to different sources of noise. Moreover, the analysis of the probability distribution function of the distance could also provide us with a better understanding of the behavior of the platoon system under different noise conditions.

In the context of the control signal plot, the values displayed represent $u(t)$, which was the control signal at each time step t. The control signal $u(t)$ was calculated based on the PID controller's output, which combined the proportional, integral, and derivative terms of the error between the desired distance and the actual distance. The purpose of the control signal is to adjust the behavior of the following vehicle in the platoon in order to minimize the error and maintain the desired inter-vehicle distance.

In the presence of packet loss or communication limitations, some control signals might be lost, which could lead to fluctuations in the actual distance as the controller tries to compensate for the missing information. The control signal plot visualizes the $u(t)$ values over time, allowing for the assessment of the system's performance and robustness in the face of communication challenges.

In this scenario, the potential impact of the secure communication between vehicles on the adaptive PID controller is considered. It should be noted that encryption is not typically used at the physical layer in vehicular networks, and the primary impact of encryption in this context would be the added delay introduced by the encryption and decryption process. However, for completeness, a demonstration of how one might implement encrypted communication using the Advanced Encryption Standard (AES) for symmetric encryption with the Python cryptography library is provided as follows.

- encrypt_data(data, key, iv): encrypts the given data using the provided key and initialization vector (IV) with the AES in CBC mode and PKCS7 padding;
- decrypt_data(encrypted_data, key, iv): decrypts the given encrypted data using the provided key and initialization vector (IV) with the AES in CBC mode and PKCS7 padding.

In the control loop, the following steps were performed:

1. Calculate the control signal as usual;
2. Encrypt the control signal using encrypt_data() with a randomly generated AES key and IV;

3. Transmit the encrypted control signal. In a real system, this would involve sending the data between vehicles;

4. Decrypt the received encrypted control signal using decrypt_data() with the same AES key and IV;

5. Continue with the decrypted control signal in the control loop.

The primary takeaway from this exercise is the potential impact of the additional latency that was introduced by the encryption and decryption process. The simulation results may not significantly differ from the unencrypted scenario, as depicted in Figures 3.15 and 3.16, as the added delay from encryption and decryption was not incorporated into this demonstration. In practice, the delay should be taken into account when analyzing the performance of the adaptive PID controller.

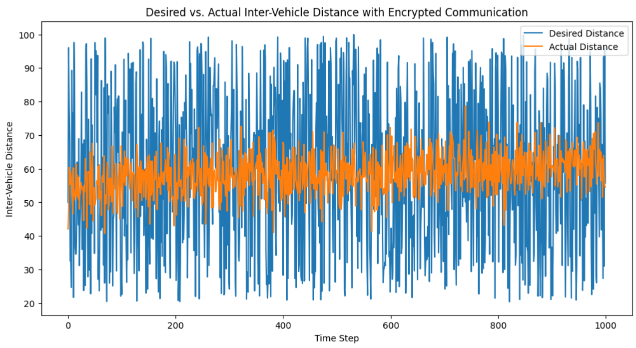


Figure 3.15: Desired vs. actual inter-vehicle distance with encrypted communication.

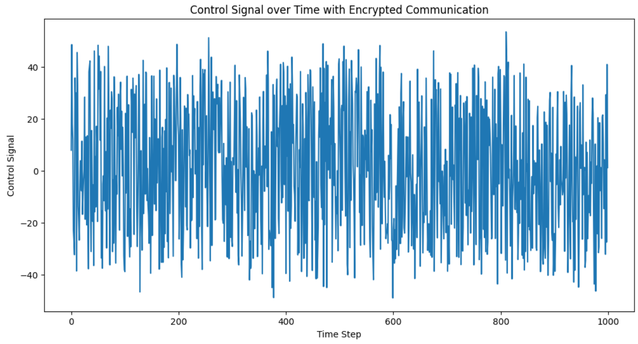


Figure 3.16: Control signal over time with encrypted communication. The control signal represents the adjustment applied to the system to maintain the desired distance between the vehicles in a truck platoon.

3.4 Integration with Large Language Models

Large Language Models (LLMs) [27, 28, 29, 30] are advanced ML models trained on vast amounts of text data. These models have achieved state-of-the-art results in various natural language understanding tasks, including text generation, translation, summarization, and question answering. LLMs are capable of understanding the context, generating coherent responses, and providing valuable insights based on the data they are exposed to. In this section, we explore the integration of LLMs into a control loop system and demonstrate their potential to enhance the system's performance.

The GPT-3.5-turbo is built on the Transformer architecture, which was first introduced in [31]. The architecture employs self-attention mechanisms [32] that enable the model to process and understand long-range dependencies in the input text. Mathematically, the self-attention mechanism can be expressed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{3.3}$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The softmax function normalizes the attention scores and helps the model to focus on the most relevant parts of the input text.

The training process involves optimizing the model's parameters in order to minimize the cross-entropy loss between the predicted token probabilities and the actual target tokens in a given context. This was carried out using the synthetic dataset that we generated. The loss function can be expressed as follows:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, \mathbf{x}; \theta), \tag{3.4}$$

where $\mathcal{L}(\theta)$ is the loss function, y_t is the target token at time step t, $\mathbf{x}$ is the input context, and θ represents the model's parameters.

During inference, the LLM generates the text by sampling from the probability distribution over the vocabulary. The model employs a temperature parameter (τ) to control the randomness of the generated text. Lower values of τ result in more deterministic outputs, whereas higher values increase the diversity of the generated text. The probability of selecting a token y_t at time step t can be defined as follows:

$$p(y_t | y_{1:t-1}, \mathbf{x}; \theta, \tau) = \frac{\exp(f_\theta(y_{1:t-1}, \mathbf{x})/\tau)}{\sum_{y't} \exp(f\theta(y'_{1:t-1}, \mathbf{x})/\tau)}, \tag{3.5}$$

where f_θ is the model's prediction function.

In the following example, we use a control loop system designed to maintain a desired inter-vehicle distance in an autonomous vehicle. The control loop's primary function is to calculate the control signals that help the vehicle maintain the desired distance. The system employs an adaptive PID controller that predicts the distance based on input data, calculates the error, and adjusts the control signals accordingly.

The control loop system operates in a noisy communication environment, simulating real-world scenarios, where input data can be affected by noise. The code snippet below defines the control loop with noise:

$$\text{input\_data\_noisy} = \text{input\_data} + \mathcal{N}(0, \text{noise\_stddev}), \tag{3.6}$$

where $\mathcal{N}(0, \text{noise\_stddev})$ represents the Gaussian noise with mean 0 and standard deviation noise_stddev.

The control loop calculates the error between the desired and predicted distances:

$$\text{error} = \text{desired\_distance} - \text{predicted\_distance}. \tag{3.7}$$

Using proportional (K_p), integral (K_i), and derivative (K_d) gains, the control signal is calculated as follows:

$$\begin{aligned} \text{control\_signal} &= K_p \cdot \text{error} + \\ K_i \cdot \text{accumulated\_error} &+ K_d \cdot \text{error\_derivative}. \end{aligned} \tag{3.8}$$

LLMs can be integrated into the control loop to provide performance updates and insights at specific intervals, as depicted in the flow diagram in Figure 3.1. In this example, the LLM is queried every 100 time steps. The LLM receives the input data string and returns a performance update.

> Time step 800: Based on the given input data, it appears that the vehicle is very close to the desired distance but still has a small positive error. The error_derivative is positive, which means that the vehicle is moving closer to the desired distance. The accumulated_error is quite high, indicating that the vehicle has been consistently either too far or too close to the desired distance over multiple readings. The vehicle_speed is higher than the speed limit, and the negative acceleration suggests that the vehicle is slowing down. The road_grade is negative, which means that the vehicle is driving uphill. Overall, the system seems to be performing adequately in that the vehicle is kept close to the desired distance, but it may need some adjustments to reduce the accumulated_error and maintain a safer speed.

> Time step 900: Based on the input data provided, it seems that the current error is positive (2.018650451683051), which indicates that the vehicle is lagging behind the desired distance. However, the error_derivative value is positive (1.0185617658178303), which indicates that the error is increasing over time. This suggests that the vehicle may be accelerating too slowly to catch up to the desired distance. The accumulated_error value is quite high (439.6535339888033), which indicates that the system has been struggling to maintain the desired distance for some time. This can be problematic in the long run as it may lead to overshooting the desired distance. The vehicle_speed value (68.55087801368882) and acceleration value (−7.270586245729461) indicate that the vehicle is currently slowing down. The road_grade value (−3.7167413781123546) suggests that the vehicle is driving uphill, which may be contributing to the slower speed. Overall, the PID system may need to adjust its control parameters to better maintain the desired distance and prevent overshooting. Additionally, the system may need to take into account the current road grade to adjust the speed and acceleration appropriately.

These updates can help engineers analyze the performance of the control loop and potentially suggest improvements or detect issues. Integrating Large Language Models into control loop systems can provide valuable insights, performance updates, and support when optimizing the system's performance. The example provided demonstrates how LLMs can be effectively used in conjunction with an adaptive PID controller in a noisy communication environment. This approach opens up new possibilities for leveraging the power of LLMs in various control applications across different domains.

3.5 Discussion

In this study, we delved into the integration of LLMs, particularly the GPT-3.5-turbo, into the control loop of a convoy of autonomous vehicles. We demonstrated the potential of LLMs to bolster the control loop performance by providing immediate feedback and recommendations. Furthermore, we scrutinized the impact of variables such as noisy communication, encryption, latency, packet loss, and communication range on the performance of the system, underlining the importance of secure and reliable communication for safety-critical applications.

An area of promising potential application for our findings is in the domain of unmanned aerial vehicles (UAVs), or drones. Similar to autonomous vehicles, drones require sophisticated control mechanisms to ensure stable flight and efficient route planning. PID controllers are a crucial component in drone flight systems as they are responsible for achieving and maintaining the drone's balance and orientation based on sensory input. The integration of LLMs could provide additional layers of interpretability and adaptability to these systems, potentially leading to safer and more reliable drone operations.

Our experiments demonstrated that the amalgamation of LLMs and conventional control techniques can indeed enhance the performance of complex systems such as autonomous vehicle platoons. LLMs provide new opportunities to leverage their natural language understanding capabilities for a plethora of applications, including diagnostics, decision making, and real-time system optimization. Our research accentuates the potential of LLM integration across an extensive range of engineering domains, where they can supplement and augment conventional control techniques.

While the use of synthetic data enabled us to demonstrate the virtues of our control scheme within a controlled environment, we acknowledge the necessity for further validation using real-world data. The focus of this chapter was the innovative combination of an adaptive PID controller with an LLM to enhance explainability. Therefore, we were primarily concentrated on the theoretical framework and its potential implementations.

3.6 Conclusions and Future Work

This study offered valuable insights into the design and implementation of AI-driven control systems for truck platooning within B5G networks and showcased the promising potential of LLMs in advanced telecommunication environments. Future work should aim to utilize actual data from real-world truck platooning systems, which could thereby provide a rigorous evaluation of our proposed control strategy.

Future research directions for LLMs could include the following:

- Assessment of forthcoming LLM architectures and the training methodologies' impact on the efficiency of adaptive control systems;
- Development of methods for the real-time fine-tuning of LLMs, allowing for swift adaptation to dynamic environments;
- Exploration of LLM applicability to other safety-critical domains such as aerospace and medical systems;
- Advancement of LLM-based control strategies in multiagent systems;

- Investigation of the combination of LLMs with alternative AI approaches, such as reinforcement learning.

For the adaptive PID control, future research could entail the following:

- Extension of the framework to cater to complex systems with nonlinear dynamics;
- Examination of the performance of different RL algorithms, such as Q-Learning or Actor–Critic, for the tuning of the PID controller gains;
- Development of hybrid control strategies that combine adaptive PID control with other control approaches;
- Integration of advanced sensing and communication technologies, such as LiDAR or V2X communication;
- Exploration of the integration of multiple control systems, such as a multiagent system, for larger-scale control problems.

Bibliography

[1] Lesch, V.; Breitbach, M.; Segata, M.; Becker, C.; Kounev, S.; Krupitzer, C. An overview on approaches for coordination of platoons. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 10049–10065. [CrossRef]

[2] Hoef, S.v.; Johansson, K.H.; Dimarogonas, D.V. Computing feasible vehicle platooning opportunities for transport assignments. *IFAC-PapersOnLine* **2016**, *49*, 43–48. [CrossRef]

[3] Boysen, N.; Briskorn, D.; Schwerdfeger, S. The identical-path truck platooning problem. *Transp. Res. Part B Methodol.* **2018**, *109*, 26–39. [CrossRef]

[4] Larsson, E.; Sennton, G.; Larson, J. The vehicle platooning problem: Computational complexity and heuristics. *Transp. Res. Part C Emerg. Technol.* **2015**, *60*, 258–277. [CrossRef]

[5] Gungor, O.E.; Al-Qadi, I.L. All for one: Centralized optimization of truck platoons to improve roadway infrastructure sustainability. *Transp. Res. Part C Emerg. Technol.* **2020**, *114*, 84–98. [CrossRef]

[6] Faber, T.; Sharma, S.; Snelder, M.; Klunder, G.; Tavasszy, L.; van Lint, H. Evaluating traffic efficiency and safety by varying truck platoon characteristics in a critical traffic situation. *Transp. Res. Rec.* **2020**, *2674*, 525–547. [CrossRef]

[7] Yang, J.; Peng, W.; Sun, C. A learning control method of automated vehicle platoon at straight path with ddpg-based pid. *Electronics* **2021**, *10*, 2580. [CrossRef]

[8] Hartenstein, H.; Laberteaux, K.P. A tutorial survey on vehicular ad hoc networks. *IEEE Commun. Mag.* **2008**, *46*, 164–171. [CrossRef]

[9] Lu, Z.; Qu, G.; Liu, Z. A survey on recent advances in vehicular network security, trust, and privacy. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 760–776. [CrossRef]

[10] Yang, Y.; Hua, K. Emerging technologies for 5G-enabled vehicular networks. *IEEE Access* **2019**, *7*, 181117–181141. [CrossRef]

[11] Chai, S.; Yin, J.; D'Ariano, A.; Samà, M.; Tang, T. Scheduling of coupled train platoons for metro networks: A passenger demand-oriented approach. *Transp. Res. Rec.* **2023**, *2677*, 1671–1689. [CrossRef]

[12] Vinitsky, E.; Kreidieh, A.; Flem, L.L.; Kheterpal, N.; Jang, K.; Wu, C.; Wu, F.; Liaw, R.; Liang, E.; Bayen, A.M. Benchmarks for reinforcement learning in mixed-autonomy traffic. In Proceedings of the 2nd Conference on Robot Learning, Zurich, Switzerland, 29–31 October 2018; Volume 87, pp. 399–409.

[13] Shladover, S. Cooperative (rather than autonomous) vehicle-highway automation systems. *IEEE Intell. Transp. Syst. Mag.* **2018**, *10*, 10–19. [CrossRef]

[14] Lu, X.-Y.; Shladover, S.E. Automated truck platoon control and field test. In *Road Vehicle Automation*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 247–261.

[15] Boubakri, A.; MataliGmmar, S. Platoons formation management strategies based on reinforcement learning. In Proceedings of the Advances in Systems Engineering: Proceedings of the 28th International Conference on Systems Engineering (ICSEng 2021), Wrocław, Poland, 14–16 December 2021; pp. 57–66.

[16] Åström, K.J.; Hägglund, T. *PID Controllers: Theory, Design, and Tuning*; Instrument Society of America: Research Triangle Park, NC, USA 1995.

[17] Åström, K.J.; Wittenmark, B. *Adaptive Control*; Dover Publications: Mineola, NY, USA, 2013.

[18] Franklin, G.F.; Powell, J.D.; Emami-Naeini, A. *Feedback Control of Dynamic Systems*; Prentice Hall: Hoboken, NJ, USA, 2014.

[19] Ghosal, A.; Sagong, S.U.; Halder, S.; Sahabandu, K.; Conti, M.; Poovendran, R.; Bushnell, L. Truck platoon security: State-of-the-art and road ahead. *Comput. Netw.* **2021**, *185*, 107658. [CrossRef]

[20] Lai, C.; Lu, R.; Zheng, D.; Shen, X. Security and privacy challenges in 5G-enabled vehicular networks. *IEEE Netw.* **2020**, *34*, 37–45. [CrossRef]

[21] Bequette, B.W. *Process Control: Modeling, Design, and Simulation*; Prentice Hall: Hoboken, NJ, USA 2018.

[22] Gu, D.-W.; Petkov, P.H.; Konstantinov, M.M. *Robust Control Design with MATLAB*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003.

[23] Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2022**, arXiv:2104.13921.

[24] Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.

[25] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[26] Ziegler, J.G.; Nichols, N.B. Optimum settings for automatic controllers. *Trans. ASME* **1942**, *64*, 759–768. [CrossRef]

[27] Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.

[28] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf` (accessed on 26 April 2023).

[29] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. 2019. Available online: `https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf` (accessed on 26 April 2023).

[30] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019 ; Volume 1, pp. 4171–4186.

[31] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.;Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008

[32] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.

Chapter 4

Emergent Cooperation and Strategy Adaptation in MASs: An EC Theory with LLMs

I. de Zarzà, J. de Curtò, Gemma Roig, Pietro Manzoni and Carlos T. Calafate. (2023). "Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs." Electronics, vol(12), 2722. DOI: 10.3390/electronics12122722

The increasing complexity of Multi-Agent Systems (MASs), coupled with the emergence of Artificial Intelligence (AI) and Large Language Models (LLMs), have highlighted significant gaps in our understanding of the behavior and interactions of diverse entities within dynamic environments. Traditional game theory approaches have often been employed in this context, but their utility is limited by the static and homogenous nature of their models. In response to this pressing need, we propose an Extended Coevolutionary (EC) Theory.

4.1 Introduction

The modern world is increasingly characterized by complex systems and interactions. These systems often involve a multitude of diverse entities, ranging from individuals and organizations to autonomous agents in Artificial Intelligence (AI)-driven environments. At the heart of understanding these complex interactions is strategic decision making, which is a vital aspect in economics, sociology, biology, and, more recently, in AI.

The study of strategic decision making has long been an essential aspect of understanding interactions among diverse entities in various domains, such as economics, sociology, and biology. Classical game theory, which was pioneered by John von Neumann and further developed by John Nash [2, 1], has provided a foundational framework for analyzing these interactions and predicting the outcomes of strategic choices. However, with the rapid advancements in AI and the emergence of Large Language Models (LLMs), there is a growing need to develop new theoretical frameworks that can better capture the dynamics of Multi-Agent Systems (MASs) in the presence of these disruptive forces [3, 4, 5, 6].

One of the key challenges in modeling strategic interactions is the inherent complexity of the environments and agents involved. In real-world scenarios, entities often have diverse characteristics, such as different risk aversions, social preferences, and learning capabilities, that can significantly influence their decision-making processes [7]. Moreover, these entities interact through various channels, including economic transactions, social relationships, and information exchange, which can further complicate the analysis of their strategic behaviors [8].

Human–Computer Interaction (HCI) is a multidisciplinary field that focuses on the design, implementation, and evaluation of interactions between humans and computers. It encompasses a wide range of topics, including the joint performance of tasks by humans and computers; the structure of communication between humans and computers; human capabilities to use computers; algorithms and programming of the interface itself; engineering concerns that arise in designing and building interfaces; the process of specification, design, and implementation of interfaces; and design trade offs.

Multi-Agent Systems (MASs) represent a paradigm in AI that models complex systems as a collection of autonomous agents that are each capable of reactive, proactive, and social behavior. These agents, which can be software programs or physical entities, interact with one another and their environment to achieve individual or shared objectives. Key concepts in MASs include coordination and control; reasoning and planning; and learning and adaptation.

In this study, we explored the intersection of HCI and MASs by integrating the EC framework with Large Language Models (LLMs) [9, 10] to model and simulate the dynamics of cooperation and defection in MASs. The EC framework combines elements from game theory, coevolutionary algorithms, and MASs to analyze and predict the behavior of agents in various interaction scenarios. By incorporating LLMs as AI agents that can provide strategic recommendations and influence human decision making, we aim to create a more comprehensive model of HCI in the context of MASs.

The core of our proposal lies in the use of intelligent sensors and sensor networks as a means to facilitate the communication and cooperation between human and intelligent agents. These sensors enable the collection of valuable data and allow for real-time adaptation and learning in response to changing environmental conditions or agent interactions. By integrating MASs and HCI, we hope to develop novel technologies and solutions centered around the use of intelligent sensors in various applications, thereby ultimately enhancing the effectiveness and efficiency of MASs in diverse HCI contexts.

HCI plays a critical role in understanding and facilitating effective cooperation between humans and intelligent agents within MASs. While HCI encompasses a wide range of topics, in this chapter, we emphasize the societal and economic perspectives of interactions between humans and AI-driven entities, such as LLMs. These perspectives involve the exchange of information, the joint performance of tasks, and the influence of AI-based strategic recommendations on human decision-making processes. By integrating HCI and MASs, we aim to create a comprehensive model that captures the evolving nature of interactions in complex systems, thereby ultimately offering insights into promoting cooperation, enhancing social welfare, and building resilience in multi-agent environments.

At the core of our proposal, we regard LLMs as intelligent sensors or AI agents that interact with human counterparts within MASs. These LLMs, which can be conceived as advanced AI-driven entities or even embodied as robots, provide strategic recommendations, process information, and influence human decision-making processes. By integrating LLMs as intelligent sensors within MASs, we facilitate the collection of valuable data that enables real-time adaptation and learning in response to changing environmental conditions or agent interactions. Our approach aims to develop novel technologies and solutions that center on the use of intelligent sensors and robots in various applications, thereby ultimately enhancing the effectiveness and efficiency of MASs across diverse HCI contexts.

Traditional approaches have largely relied on game theory. However, as the digital era progresses, disruptive forces such as AI and LLMs are transforming the landscape of strategic decision making. These advancements underline the pressing need for new theoretical frameworks that are capable of capturing the nuanced dynamics of MASs amidst this transformative wave.

To this end, we introduce an Extended Coevolutionary (EC) Theory as an alternative to traditional game theory approaches for modeling and analyzing strategic interactions among heterogeneous agents. Our EC framework aims to capture the evolving nature of MASs and incorporate the potentially disruptive influence of LLMs on business and society. The main contributions of this study are:

1. The development of a comprehensive theoretical framework that integrates coevolutionary dynamics, adaptive learning, and LLM-based strategy recommendations for understanding the emergence of cooperation and defection patterns in MASs.
2. The design of a simulation environment that allows for the exploration of the EC framework, thus incorporating heterogeneous agents and multi-layer networks to model diverse interactions among entities.
3. The evaluation of the effectiveness of the EC framework in promoting cooperative behavior and robustness in the face of disruptions by using various performance metrics and advanced visualization techniques.

By achieving these objectives, we hope to provide valuable insights into the interplay between strategic decision making, adaptive learning, and LLM-informed guidance in complex, evolving systems. Our findings have the potential to inform the development of novel strategies and interventions for harnessing the power of AI and LLMs in promoting cooperation, enhancing social welfare, and building resilience in multi-agent environments.

The remainder of this chapter is organized as follows: In Section 4.2, we provide a comprehensive review of the related work that covers topics such as game theory and NASH equilibrium, coevolutionary algorithms, MASs, and AI. Section 4.3 presents the EC framework and discusses its key components, such as coevolutionary dynamics, adaptive learning, and the role of LLMs in strategy formation. Section 4.3.4 introduces the concept of LLMs in the EC framework and explains how they can be used to generate strategy recommendations and influence agent interactions. In Section 4.4, we present the methodology, which provides proofs of the EC framework to establish its mathematical foundations. Section 4.5 details the simulation environment used in our experiments, including implementation details, performance metrics, and visualization techniques. Section 5.5 presents the results and analysis of our experiments by examining the emergence of cooperation and defection patterns, the influence of LLM-based strategy recommendations, and the overall system robustness and resilience. Section 4.7 discusses the broader implications of our findings for business and society, as well as the limitations of our current framework and potential avenues for future work. Finally, Section 4.8 concludes the chapter by summarizing our key findings and contributions to the field of MASs and HCI.

4.2 Related Work and Theoretical Context

Game theory is a mathematical framework for studying strategic interactions among rational agents [11]. A central concept in game theory is the NASH equilibrium, which is a state in which no player can improve their utility by unilaterally changing their strategy, given the strategies of the other players [1]. The concept of NASH equilibrium has been widely applied to model and analyze a variety of strategic situations, including economic transactions, social dilemmas, and political negotiations [12]. Recent research has explored the extensions of classical game theory to incorporate more realistic assumptions about agent behavior and the dynamics of strategic interactions, such as bounded rationality, learning, and adaptation [13, 14]. These extensions have led to the development of new solution concepts and methods for predicting and influencing the outcomes of strategic interactions in complex, evolving environments.

Coevolutionary algorithms are a class of evolutionary algorithms that model the adaptive processes of learning and optimization in populations of interacting agents [15]. In coevolutionary algorithms, agents adapt their strategies over time in response to the strategies of other agents in the population, thereby leading to the emergence of complex patterns of cooperation, competition, and specialization [16, 17, 18]. These algorithms have been used to study a wide range of problems in AI, optimization, and MASs, including the evolution of cooperation in social dilemmas [19, 20, 21], the development of efficient algorithms for hard optimization problems [22, 23], and the emergence of communication and coordination in MASs [24, 25].

MASs [26] are part of a subfield of AI that focuses on the development of computational models and algorithms for simulating and controlling the interactions among multiple autonomous agents [27, 28, 29]. MAS research aims to understand the underlying principles that govern the behavior of complex, distributed systems, and to develop methods for coordinating the actions of individual agents to achieve global objectives [30, 31].

Recent advances in AI, particularly in the areas of machine learning and Large Language Models (LLMs) [32, 33, 34], have opened up new possibilities for modeling and analyzing strategic interactions in MASs [35]. While there is limited research on the direct integration of LLMs in this specific setting, our work aims to bridge this gap and explore the potential impacts of AI on the dynamics of cooperation, competition, and social welfare in evolving multi-agent environments. The infusion of LLM-based advice into agent decision making opens up promising avenues for investigation, particularly regarding the potential benefits and challenges posed by AI-driven guidance in MASs. Notably, LLMs, such as GPT-3.5-turbo, are capable of generating human-like natural language text, thereby allowing them to provide strategic guidance and recommendations to agents in a Multi-Agent System [36]. By incorporating LLM-based advice into the decision-making processes of agents, researchers have begun investigating the potential benefits and challenges that may arise from AI-driven guidance in MASs.

For instance, recent studies have shown that LLMs can enhance the performance of agents in various tasks, such as negotiation [37] and coordination [38], by providing real-time strategic recommendations based on the current state of the environment and agent interactions. These initial findings suggest that LLMs can play a significant role in shaping the dynamics of multi-agent systems and, ultimately, the outcomes of strategic interactions.

In summary, while the direct integration of LLMs in the context of strategic interactions and MASs is still an emerging area of research, our work aims to contribute to the understanding of the potential benefits and challenges associated with incorporating AI-driven guidance in complex, evolving environments. By extending existing theories and methodologies, such as coevolutionary algorithms and game theory, our proposed Extended Coevolutionary (EC) framework seeks to capture the unique characteristics of LLMs and their potential impact on the dynamics of cooperation, competition, and social welfare in Multi-Agent Systems.

4.3 EC Theory

In this section, we present the Extended Coevolutionary (EC) Theory framework, which is the main contribution of our work. Our EC framework integrates concepts from game theory, coevolutionary algorithms, and AI to study the emergence and evolution of cooperation and defection in Multi-Agent Systems (MASs). Specifically, the EC framework extends classical game-theoretic models [17, 39] by incorporating adaptive learning, heterogeneous agents, and multi-layer network structures. Moreover, we introduce the use of LLMs, such as GPT-3.5-turbo, to assist agents in forming their strategies, thereby enabling a more comprehensive understanding of the dynamics of strategic interactions in complex environments. In the following subsections, we detail the key components and theoretical tools used in the development of the EC framework.

4.3.1 Coevolutionary Dynamics and Adaptive Learning

Coevolutionary dynamics are central to our proposed EC framework, as they capture the process by which agents adapt their strategies in response to the strategies of others in the population. The EC framework employs adaptive learning mechanisms in which agents update their strategies based on the utilities they receive from interacting with other agents.

Let s_z denote the strategy of agent z, and let $U_z(s_z, s_{-z})$ represent the utility of agent z given its own strategy s_z and the strategies of all other agents s_{-z}. The adaptive learning process can be described by the following update rule:

$$s_z(t+1) = s_z(t) + \alpha \nabla U_z(s_z(t), s_{-z}(t)), \tag{4.1}$$

where α is the learning rate, and ∇U_z is the gradient of the utility function with respect to the strategy s_z. This update rule captures the process by which agents adjust their strategies to maximize their utilities based on the current state of the population.

4.3.2 Large Language Models in Strategy Formation

In our EC framework, we also incorporated the use of LLMs, such as GPT-3.5-turbo, to assist agents in forming their strategies. These AI agents can provide valuable insights and recommendations based on the current state of the game and the strategies of neighboring agents. By integrating LLMs into the adaptive learning process, we can explore how the introduction of AI agents influences the dynamics of cooperation and defection in MASs; an illustrative diagram can be seen in Figure 4.1.

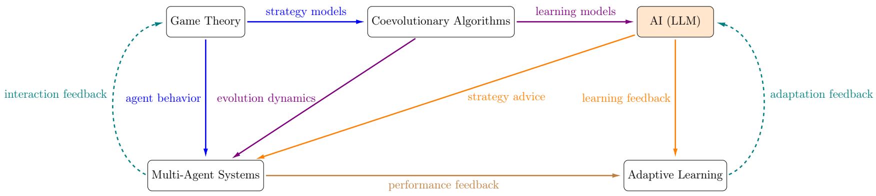


Figure 4.1: A detailed schematic representation of the Extended Coevolutionary (EC) Theory framework emphasizing the integration of Large Language Models (LLMs). The diagram not only illustrates the primary components of the framework—game theory, coevolutionary algorithms, AI (LLM), Multi-Agent Systems, and adaptive learning—but also explicates their dynamic interconnections. Feedback loops are introduced to signify ongoing adaptation and learning processes, while labeled arrows illuminate the nature of interactions, such as strategy modeling, agent behavior, learning models, evolution dynamics, strategy advice, learning feedback, and performance feedback. This comprehensive portrayal seeks to foster a deeper understanding of the intricate dynamics within the EC framework.

Indeed, the feedback process in adaptive learning extends beyond modifying interaction strategies. Adaptive learning involves an iterative process of adjusting the model parameters based on the feedback received, thereby continuously improving the performance of the model. In our EC framework, adaptive learning not only informs the strategies adopted by agents, but also refines the underlying models that drive agent behavior. Specifically, the "learning feedback" from the LLM to the adaptive learning component of the system captures this process of continuous improvement. When the LLM provides strategic advice to the agents, it includes not only immediate actions, but also feedback on the current strategies. This feedback is then used to adjust the models that inform agent behavior, thereby enabling them to learn and adapt over time. Moreover, the "adaptation feedback" from the adaptive learning component back to the LLM signifies the updates in model parameters based on the performance and interaction feedback. This continuous feedback loop ensures that the LLM, and, thus, the strategies it recommends, evolves over time to better support agent interactions.

4.3.3 Heterogeneous Agents and Multi-Layer Network Model

The EC framework acknowledges the importance of agent heterogeneity and complex network structures in shaping the dynamics of strategic interactions. We modeled agents with varying characteristics, such as different levels of risk aversion, social preferences, and learning capabilities. Furthermore, we introduced a multi-layer network model that captures multiple types of interactions between agents, such as economic transactions, social relationships, and information exchange.

The multi-layer network is represented by a tuple $G = (V, E_1, E_2, \ldots, E_k)$, where V is the set of nodes (agents), and E_z is the set of edges (interactions) in layer z. The multi-layer network allows us to study the interdependencies between different types of interactions and their effects on cooperation and defection dynamics in the population.

4.3.4 Large Language Models in EC

LLMs, such as GPT-3.5-turbo, play a significant role in the EC framework, especially in the context of strategic formation and adaptive learning. These AI agents can analyze the current state of the game, the strategies employed by neighboring agents, and can provide valuable insights and recommendations for the agents' next actions. The incorporation of LLMs within the EC framework enables a deeper understanding of the dynamics of cooperation and defection, as well as the influence of AI agents on the overall system.

LLM-Based Adaptive Learning

In the EC framework, LLMs are used to support agents during the adaptive learning process. At certain intervals, agents consult the LLM for advice on their next strategic move while considering the strategies of their neighbors. To formalize this interaction, let $Q_{z,t}$ be the LLM's recommendation for agent z at time t. We can express the recommendation as a function of the neighboring agents' strategies $s_{-z}(t)$:

$$Q_{z,t} = f(s_{-z}(t)), \tag{4.2}$$

where $f(\cdot)$ is the function representing the LLM's recommendation process.

In the context of real-time applications, the function $f(\cdot)$ needs to be efficient and robust. Efficiency is required to ensure that the recommendation process does not introduce significant latency into the system, which is especially critical in real-time applications where timely response is often necessary. Robustness, on the other hand, is needed to ensure that the recommendation process can handle a wide range of possible inputs and still produce meaningful outputs. This is crucial in a dynamic Multi-Agent System where the strategies of neighboring agents can vary significantly over time. In the context of LLMs, the function $f(\cdot)$ is implemented by the LLM's underlying ML model. The model is trained on a large corpus of data and is capable of generating strategic recommendations based on the input it receives. The specifics of this process depend on the architecture and training of the LLM. In the case of GPT-3.5-turbo, for example, the model takes the current context, including the strategies of neighboring agents, and generates a recommendation based on patterns it has learned during its training.

The agent's strategy update can then be modeled as a combination of its original adaptive learning process and the LLM's recommendation:

$$s_z(t+1) = (1-\beta)\left(s_z(t) + \alpha \nabla U_z(s_z(t), s_{-z}(t))\right) + \beta Q_{z,t}, \tag{4.3}$$

where $\beta \in [0,1]$ represents the influence rate of the LLM on the agent's strategy. When $\beta = 0$, the agent relies solely on its original adaptive learning process; when $\beta = 1$, the agent fully adopts the LLM's recommendation.

Incorporating LLM Uncertainty

Given the probabilistic nature of LLM-generated recommendations, it is essential to consider the uncertainty associated with the LLM's advice. One way to account for this uncertainty is to introduce a confidence measure $c_{z,t}$ that is associated with the LLM's recommendation:

$$c_{z,t} = g(Q_{z,t}), \tag{4.4}$$

where $g(\cdot)$ is a function that maps the LLM's recommendation to a confidence value in the range $[0,1]$.

By incorporating the confidence measure, we can adjust the agent's strategy update rule as follows:

$$s_z(t+1) = (1 - \beta c_{z,t})\left(s_z(t) + \alpha \nabla U_z(s_z(t), s_{-z}(t))\right) + \beta c_{z,t} Q_{z,t}. \tag{4.5}$$

This modified update rule allows agents to weigh the LLM's advice based on the confidence associated with the recommendation, thereby leading to a more nuanced adaptive learning process.

In summary, our EC Theory framework provides a powerful and flexible approach for studying the emergence and evolution of cooperation and defection in MASs. By incorporating adaptive learning, heterogeneous agents, multilayer network structures, and LLMs, the EC framework can offer novel insights into the complex dynamics of strategic interactions in diverse settings. Futhermore, the integration of LLMs within the EC framework provides a novel perspective on the dynamics of cooperation and defection in MASs. The LLM-assisted adaptive learning process, along with the consideration of LLM uncertainty, contributes to a more comprehensive understanding of the complex strategic interactions in diverse settings.

4.4 Methodology

Let us assume that our EC framework can be reduced to a simple two-player game with finite strategy sets and that the utility functions incorporate only the immediate payoffs without the adaptive learning mechanisms or LLM-based strategy recommendations. The proof below demonstrates the existence of a NASH equilibrium for this simplified game using Brouwer's fixed-point theorem.

Theorem 1. *Given a two-player game in the EC framework with each player having a finite set of strategies and where the utility functions are based only on immediate payoffs, without any adaptive learning mechanisms or Large Language Model (LLM) based strategy recommendations, there exists a NASH equilibrium.*

Proof. Let us consider a two-player game represented by the EC framework, with each player o having a finite set of strategies S_o, where $o \in \{1, 2\}$. Let $\mathbf{s} = (s_1, s_2)$ denote a strategy profile, where $s_o \in S_o$ for both players.

1. Define the utility functions $u_o(\mathbf{s})$ for each player o as the immediate payoffs from the chosen strategy profile $\mathbf{s}$.

2. Define the best response correspondence $B_o : S_{-o} \to S_o$ for each player o, which maps a strategy of the opponent to the set of best responses for player o. Since S_o is finite, the best response correspondence is nonempty and upper hemicontinuous.

3. Define the correspondence
$G : S_1 \times S_2 \to S_1 \times S_2$ as $G(s_1, s_2) = (B_1(s_2), B_2(s_1))$. This maps a strategy profile $\mathbf{s}$ to the set of best response profiles for both players. Since B_o is nonempty and upper hemicontinuous for both players, G is also nonempty and upper hemicontinuous.

4. Define the strategy space $S = S_1 \times S_2$ and assume it is a compact and convex set. Compactness follows from the finiteness of the strategy sets, and convexity follows, since we can treat the strategies as probability distributions over the pure strategies.

5. Apply Brouwer's fixed-point theorem, which states that every continuous function from a compact, convex set to itself has a fixed point. Since G is nonempty, upper hemicontinuous, and maps S to itself, it has a fixed point $\mathbf{s}^* = (s_1^*, s_2^*) \in S$.

6. At this fixed point $\mathbf{s}^*$, we have $s_1^* \in B_1(s_2^*)$ and $s_2^* \in B_2(s_1^*)$. This means that, given the strategy of the opponent, each player is choosing their best response, thus making $\mathbf{s}^*$ a NASH equilibrium.

By following these steps and applying Brouwer's fixed-point theorem, we have proven the existence of a NASH equilibrium for a simplified two-player game within the EC framework. □

An interesting point to note here is that the convex combination in (4) is proposed under the assumption that the weightings of the adaptive learning mechanism and the LLM's recommendation sum to one, which is often a mathematical convenience that helps to maintain the strategy within a defined strategy space. This is particularly important when strategies are represented as probability distributions over a finite set of pure strategies, where the sum of probabilities must equal to one. A convex combination ensures that the resulting strategy is a valid probability distribution. However, considering an affine combination could also bring an interesting perspective. An affine combination could potentially allow for a greater range of weightings and, thus, may offer more flexibility. It could provide a richer representation of how the agent might incorporate the advice from the LLM or the learning mechanism in its decision-making process. But it is important to note that using an affine combination could lead to situations where the strategy might fall outside the original strategy space, especially if the strategies are represented as probability distributions. We could indeed modify the model to allow for affine combinations of the adaptive learning mechanism and the LLM's recommendation, provided that we adjust the strategy space and the interpretation of the strategies accordingly. We could also explore different mechanisms to determine the relative weightings of the two components, beyond a simple fixed weight. For instance, the weightings could depend on the agent's confidence in the LLM's recommendation or on the performance of the adaptive learning mechanism.

Another important issue to consider is that the existing formulation does not characterize the LLM-related uncertainty and seems to be more related to the sensitivity of the agents' strategies to the LLM's recommendations. To address this point, we could propose to revise the model to explicitly consider the uncertainty in the LLM's recommendation. The LLM's recommendation $Q_{z,t}$ could be modeled as a random variable instead of a deterministic function of the neighboring agents' strategies $s_{-z}(t)$. This could better represent the inherent uncertainty of AI systems. We could also explore ways to quantify this uncertainty. For instance, we could explore this by incorporating a measure of the variance or entropy associated with the LLM's recommendation. We might also consider modifying the utility functions to reflect the agents' risk attitudes towards the LLM-related uncertainty. For example, risk-averse agents might prefer strategies that minimize the potential negative impact of an inaccurate LLM recommendation, while risk-neutral agents might be indifferent to this uncertainty. However, it is important to note that introducing uncertainty into the model may complicate the analysis. The existence of a NASH equilibrium, as demonstrated in the proof using Brouwer's fixed-point theorem, may no longer be guaranteed. This is because the fixed-point theorem assumes that the function (in this case, the correspondence G) is deterministic, whereas introducing uncertainty into the LLM's recommendation might render G stochastic.

Given the complexity of LLMs and the inherent difficulties in mathematically formalizing their properties, proving a specific aspect of the EC framework that incorporates LLM-based strategy recommendations is challenging. However, we can attempt to provide a simple proof that demonstrates the potential improvement in utility for an agent following LLM-based strategy recommendations.

The assumptions and simplifications include the following:

1. Consider a two-player game represented by the EC framework, with each player o having a finite set of strategies S_o, where $o \in \{1, 2\}$.
2. Assume that the LLM provides strategy recommendations for player 1.
3. Let the true utility functions $u_o(\mathbf{s})$ for each player o be known and fixed.
4. Assume that the LLM's recommendations are based on the true utility functions of both players and that the LLM generates recommendations that maximize player 1's expected utility, given player 2's strategy.

The proof indeed builds on several strong assumptions and simplifications, especially the third one, where we assume that the true utility functions are known and fixed. This is, of course, an oversimplification; in real-world scenarios, utility functions might be unknown or dynamically changing. This assumption is made primarily to make the proof tractable, thereby providing a simplified demonstration of the potential benefits of incorporating LLM-based strategy recommendations. The third assumption can be interpreted as a "perfect information" assumption. We are assuming that the LLM is omniscient and has complete information about the utility functions of both players.

Theorem 2. *Consider a two-player game represented by the EC framework, where each player o has a finite set of strategies S_o ($o \in \{1, 2\}$), and the true utility functions $u_o(\mathbf{s})$ for each player o are known and fixed. Assume that the LLM provides strategy recommendations for player 1 and that these recommendations are based on the true utility functions of both players. If the LLM's recommendations aim to maximize player 1's expected utility given player 2's strategy, then player 1's expected utility following the LLM's recommendations will be at least as high as when choosing any other strategy from their strategy set.*

Proof. Let $\mathbf{s} = (s_1, s_2)$ denote a strategy profile, where $s_o \in S_o$ for both players. Define s_1^R as the strategy recommendation provided by the LLM for player 1, given player 2's strategy s_2.

1. Define the expected utility for player 1 when following the LLM's recommendation as $E[u_1(s_1^R, s_2)]$.
2. Since the LLM generates recommendations based on the true utility functions of both players and aims to maximize player 1's expected utility, we have $E[u_1(s_1^R, s_2)] \geq E[u_1(s_1, s_2)]$ for any $s_1 \in S_1$.

3. If player 1 chooses to follow the LLM's strategy recommendation s_1^R, their expected utility will be at least as high as when choosing any other strategy from their strategy set.

□

In this simplified proof, we have shown that following LLM-based strategy recommendations can potentially improve the expected utility for player 1. However, it is important to note that this proof is built on several assumptions and simplifications that may not hold in more complex scenarios or when considering adaptive learning mechanisms and heterogeneous agents.

These proofs provide a strong foundation for understanding the theoretical aspects of the EC framework and the potential benefits of incorporating LLM-based strategy recommendations in Multi-Agent Systems.

4.5 Simulation Environment

The EC framework was implemented as a simulation environment to explore the interactions between heterogeneous agents in a multi-layer network. The simulation consists of a discrete-time system with the following steps:

1. Initialization: Create a set of N heterogeneous agents with varying characteristics such as risk aversion, social preferences, and learning capabilities. Generate a multi-layer network representing various types of interactions between agents, such as economic transactions, social relationships, and information exchange.
2. Iteration: For each time step $t \in \{1, 2, \ldots, T\}$, where T is the total number of simulation rounds:
   (a) Simulate interactions between agents based on their current strategies and update their utilities.
   (b) Apply adaptive learning to update the agents' strategies, with LLM consultations at specified intervals.
   (c) Update the network structure based on the evolving strategies and utilities
   of agents.

3. Analysis: Evaluate the system's performance using various metrics and visualize the network's evolution to gain insights into the dynamics of cooperation and defection.

The multi-layer network structure we use in the simulation is not only a complex system composed of three interconnected layers—economic', social', and 'information'—but it is also a reflection of real-world multi-agent systems. Each layer represents a distinct type of interaction among the agents. These interactions are not isolated; instead, they collectively influence the decision-making process of the agents in a holistic manner. For example, an agent's economic decisions may be influenced by their social interactions and the information they receive. Moreover, these interactions and their consequences can feedback into each layer, thereby causing changes that can further influence the decision-making process. In addition to interacting within and across layers, the agents themselves are characterized by their strategies and attributes that were previously presented. For instance, the strategies formulated in the context of the EC framework are implemented by the agents as they interact within and across the layers of the multi-layer network.

Formally, the multi-layer network can be defined as a tuple $G = (V, E_1, E_2, E_3)$, where:

- V is the set of nodes (entities) in the network, each characterized by a strategy, risk aversion, social preference, learning capability, and utility.
- E_1 represents the set of edges in the 'economic' layer indicating economic interactions between the entities.
- E_2 represents the set of edges in the 'social' layer indicating social interactions between the entities.
- E_3 represents the set of edges in the 'information' layer indicating information exchange between the entities.

The multi-layer network was constructed as a multi-graph in order to allow for multiple edges between a pair of nodes that are each associated with a different layer. The edges within each layer were generated using a random graph model with a specified edge probability. This model ensured that the network structure exhibited a random distribution of edges, thus capturing the inherent uncertainty and complexity of real-world interaction patterns among agents. The multi-layer network structure served as a robust and versatile framework for simulating the interplay of various interaction types among agents, thereby facilitating a comprehensive understanding of the system's dynamics and evolution.

While this multi-layer network structure was used here for simulation purposes, its design is representative of the type of complex multi-agent systems seen in real-world situations. By using such a structure, we can capture and study the interplay of various interaction types among agents, which is crucial for understanding the dynamics and evolution of Multi-Agent Systems.

4.5.1 Performance Metrics

To measure the effectiveness of the EC framework, several performance metrics were introduced, including overall social welfare, the prevalence of cooperation, and the robustness of the system to shocks or disruptions. The selection of these metrics was motivated by their ubiquity in Multi-Agent Systems literature and their relevance to the specific aspects we aimed to enhance through the EC framework.

- Overall social welfare: The sum of all agents' utilities at time t. This metric is traditionally used in economics and game theory to measure the total benefit accrued by all members of a system, thus providing an aggregate measure of system performance. Higher social welfare indicates that more agents are achieving higher utility, which aligns with the goal of our EC framework to improve individual and collective outcomes.

$$W(t) = \sum_{z=1}^{N} U_z(t). \tag{4.6}$$

- Prevalence of cooperation: The proportion of agents employing a cooperative strategy at time t. This metric is particularly relevant for Multi-Agent Systems where cooperative behavior can lead to mutual benefit or improved social welfare. As the EC framework aims to encourage cooperative behavior, monitoring the prevalence of cooperation provides a direct measure of this aspect of the system's performance.

$$P_c(t) = \frac{\sum_{z=1}^{N} I[s_z(t) = \text{cooperate}]}{N}, \tag{4.7}$$

where $I[\cdot]$ is the indicator function, which equals 1 if the condition inside the brackets is true and equals 0 otherwise.

- Robustness: The ability of the system to maintain cooperation levels in the face of shocks or disruptions. In Multi-Agent Systems literature, the robustness of a system is often a critical measure of performance, and it indicates how well the system can adapt to changes or uncertainties. Given that real-world Multi-Agent Systems often face dynamic environments and perturbations, we incorporated this metric to evaluate how well the EC framework could maintain performance under such conditions.

$$R(t) = \frac{P_c(t) - P_c(t-1)}{P_c(t-1)}. \tag{4.8}$$

4.5.2 Visualization Techniques

Effective visualization techniques are essential for understanding the complex dynamics of the EC framework. Several approaches can be employed to illustrate the evolution of the system, including:

1. Time-lapse network visualization: Display the network's evolution over time in an animation in order to highlight changes in network structure, agent strategies, and cooperation levels. This visualization can be created using libraries such as NetworkX or Gephi, where nodes represent agents, and edges represent relationships. The nodes' colors and sizes can be adjusted based on the cooperation levels, thereby allowing observers to track the development of cooperation and defection strategies over time.

2. Interactive visualizations: Develop interactive visualizations that allow users to explore the relationships between agents, their strategies, and the various types of interactions in the multi-layer network. This can be achieved using web-based visualization libraries such as D3.js or Plotly, which enable the creation of dynamic, responsive visualizations. For example, users could filter agents based on certain attributes, adjust time scales, or zoom into specific areas of the network to investigate local dynamics. Tooltips can also be added to display additional information about individual agents and their strategies by hovering or clicking.

3. Heatmaps: Generate heatmaps to visualize the spatial distribution of cooperation and defection strategies, thus providing insights into the emergence of clusters or patterns within the network. This can be done us-

ing Python libraries such as Matplotlib or Seaborn, where the X-axis represents rounds, the Y-axis represents agents, and the color intensity indicates the cooperation level of each agent. Such heatmaps can help identify regions of high cooperation or defection, as well as detect sudden shifts in strategies or the formation of stable cooperation clusters over time.

These visualization techniques, along with the performance metrics, provide valuable tools for analyzing the behavior of agents and the overall dynamics of cooperation and defection within the EC framework.

4.6 Results and Analysis

In this section, we present the results and analysis of our experiments with the EC framework, with a focus on the emergence of cooperation and defection patterns among heterogeneous agents in the Multi-Agent System. We investigated the role of adaptive learning and the impact of LLM-based strategy recommendations on these patterns, as well as the network's overall robustness and resilience. The network's initial and final structures, as shown in Figures 4.2 and 4.3, provide a visual representation of the evolution of these patterns over time.

The experiments conducted in this study were executed using a custom-built Multi-Agent System simulator, which was designed specifically to study the emergence of cooperation and defection patterns in complex networks. This simulator allows for the creation and manipulation of heterogeneous agents by implementing adaptive learning processes and incorporating LLM-based strategy recommendations. It is capable of simulating dynamic, evolving multi-layer networks while tracking and visualizing changes in the system over time. The simulator provides a comprehensive platform for observing and analyzing the effects of various hyperparameters and network structures on agent behavior and overall system performance. The visualizations generated by the simulator facilitate a deeper understanding of the complex dynamics at play within the Multi-Agent System, thereby enabling researchers to fine-tune the EC framework and optimize its potential for fostering cooperation in diverse real-world applications.

Through the implementation of the EC framework, we observed the emergence of cooperation and defection patterns within the Multi-Agent System. The adaptive learning process, combined with the varying characteristics of heterogeneous agents, led to the formation of clusters of cooperators and defectors within the network. These clusters evolved dynamically over time, having been influenced by the agents' strategies and interactions with their neighbors. Next, we will delve deeper into the factors contributing to these patterns and their significance in the context of the EC framework.

In the simulations conducted, a set of hyperparameters was used to determine the behavior of the agents and the network. The total number of agents, or entities, in the network was set to 100. The simulation was run over 500 rounds to observe the evolution of agent strategies and network properties. The initial cooperation factor was set to 0.5, meaning that 50% of the agents started with a cooperative strategy. The learning rate was set at 0.1, which determined the probability of agents adapting their strategies based on their neighbors' performance. To model the addition of new connections between agents, an edge addition probability of 0.05 was used, thereby allowing the network to evolve over time. Finally, a cooperation threshold of 0.6 was implemented, which represented the minimum proportion of cooperative neighbors needed for an agent to switch to a cooperative strategy. These hyperparameters guided the simulation and influenced the outcomes of social welfare and cooperation prevalence within the network.

The prevalence of cooperation increased when agents were able to learn from their neighbors' strategies, particularly in the presence of high levels of trust and reciprocity. Conversely, when agents were more risk-averse or selfish, defection patterns emerged, leading to suboptimal outcomes for both the individual agents and the system as a whole.

The incorporation of LLMs into the EC framework significantly impacted the adaptive learning process and the formation of cooperation and defection patterns, as shown in Figures 4.2–4.5. The LLM consult interval served as an effective mechanism to analyze their influence on the system. By consulting the LLMs at specific intervals, we could observe the impact of their recommendations on the agents' decision making, as well as the resulting cooperation and defection patterns over time.

When agents consulted LLMs for strategy recommendations, they were more likely to make informed decisions based on the broader context of their neighbors' strategies and the network structure. The LLM-based recommendations promoted cooperation, especially when the majority of neighbors were already cooperating, as agents sought to maximize their utilities through mutual cooperation.

The choice of using an LLM consult interval, rather than a direct comparison of the system with and without LLMs, allowed us to better understand the dynamic interplay between LLM-guided decision making and the agents' autonomous adaptive learning. This approach offers insights into the complex, evolving relationships between agents, their strategies, and the network structure, which might be obscured in a direct comparison scenario.

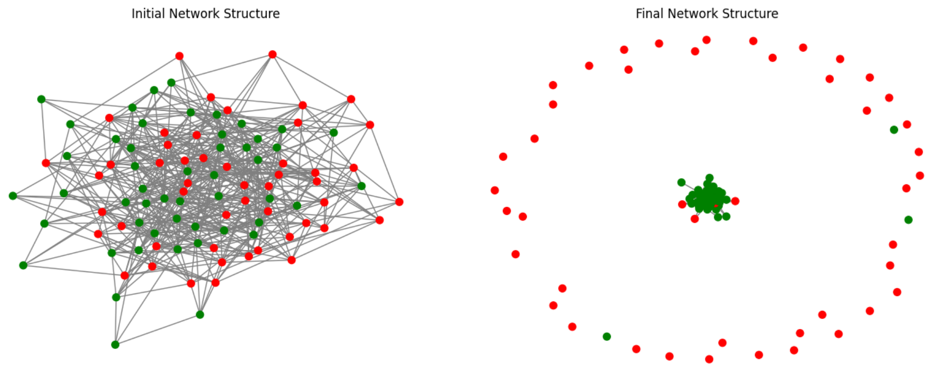


Figure 4.2: Evolution of network structure over time, illustrating the changes in cooperation and defection patterns among agents. The initial network structure (**left**) is compared to the final network structure (**right**) after running the simulation with adaptive learning, including LLM-based strategy recommendations every 10,000 rounds. The nodes are colored green if the entity's strategy is to cooperate, and red if the entity's strategy is to defect. We used a preferential attachment rule for edge creation and an edge removal rule based on a cooperation threshold of 0.6.

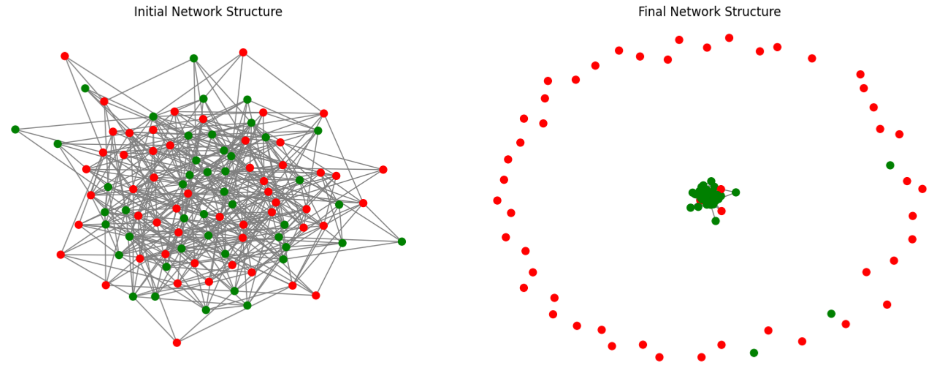


Figure 4.3: Evolution of the network structure and agent strategies, taking into account individual risk aversion, social preference, and learning capability during the simulation. The initial network structure (**left**) and the final network structure (**right**) are presented after incorporating adaptive learning and LLM-based strategy recommendations every 10,000 rounds. The nodes are colored green if the entity's strategy is to cooperate, and red if the entity's strategy is to defect. We used a preferential attachment rule for edge creation and an edge removal rule based on a cooperation threshold of 0.6. The changes in cooperation and defection patterns among agents, influenced by their unique risk aversion, social preference, and learning capability, can be observed over the course of the simulation.

Moreover, LLMs helped agents to adapt more quickly to changes in their environment, such as the emergence of defectors or fluctuations in the levels of trust and reciprocity within the network. This increased adaptability allowed the agents to maintain cooperation levels and achieve higher overall social welfare. For instance, Figure 4.4 illustrates the multi-layer network structures before and after the simulation, where they achieved an overall social welfare of 2442.3 and a prevalence of cooperation of 63.00%, with an LLM consult interval of 15,000 rounds.

The EC framework demonstrated robustness and resilience in the face of shocks and disruptions, such as the introduction of defectors or changes in the network structure. The adaptive learning process, along with the influence of LLM-based strategy recommendations, allowed agents to swiftly adjust their strategies in response to these perturbations.

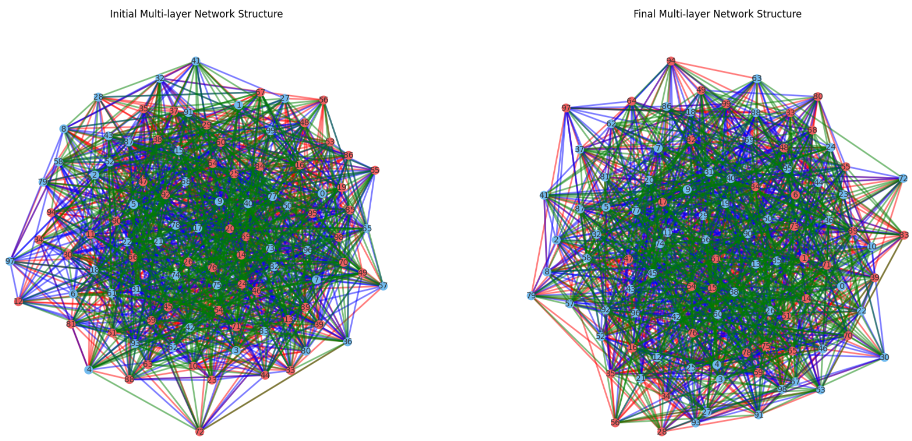


Figure 4.4: Multi-layer network structures before and after the simulation: These side-by-side plots show the multi-layer network consisting of economic (red edges), social (blue edges), and information (green edges) layers. Each layer in the network represents a different type of interaction: economic transactions, social relationships, and information exchange. The left plot represents the initial network structure, while the right plot displays the final network structure after the simulation. Nodes are colored based on their strategies, with blue representing cooperation and red representing defection. The evolution of strategies can be observed as a result of the agents' interactions, learning capabilities, and LLM-based strategy recommendations.

The system's robustness was further enhanced by the multi-layer network model, which captured different types of interactions between agents. This multi-layer structure allowed agents to maintain cooperation levels in one layer, even when facing disruptions in another layer. For instance, in the multi-layer system studied in Figure 4.4, the EC achieved a change in social welfare after a shock of 1819 and a change in cooperation prevalence after the shock of 5.00%, with an LLM consult interval of 30,000 rounds. Overall, the EC framework proved to be a resilient approach to modeling and promoting cooperation in complex MASs.

The visualizations generated during the simulation provided valuable insights into the dynamics of the EC framework. Time-lapse network visualizations revealed the emergence of cooperation and defection patterns, as well as the evolution of the network structure over time. Interactive visualizations allowed for the exploration of agent strategies, network layers, and the relationships between agents in greater detail.

As shown in Figure 4.5, the evolution of cooperation in the multi-layer network is illustrated across representative rounds. Each plot presents the state of the network at different points in time, with blue nodes representing cooperative entities and red nodes symbolizing defecting entities. Node numbers represent the unique identifiers for each agent. The cooperative prevalence values, indicated in the subcaptions, provide insights into the percentage of cooperative agents within the network at each round.

Over the course of the simulation, we can observe shifts in the prevalence of cooperation and defection within the network, as well as the formation of clusters of cooperative and defecting agents. These changes can be attributed to the adaptive learning processes, the interactions between entities across multiple layers, and the influence of LLM-based strategy recommendations. The figure provides valuable insights into the dynamics of cooperation in complex multi-layer networks and highlights the significance of considering multiple dimensions of interaction when studying the evolution of cooperation.

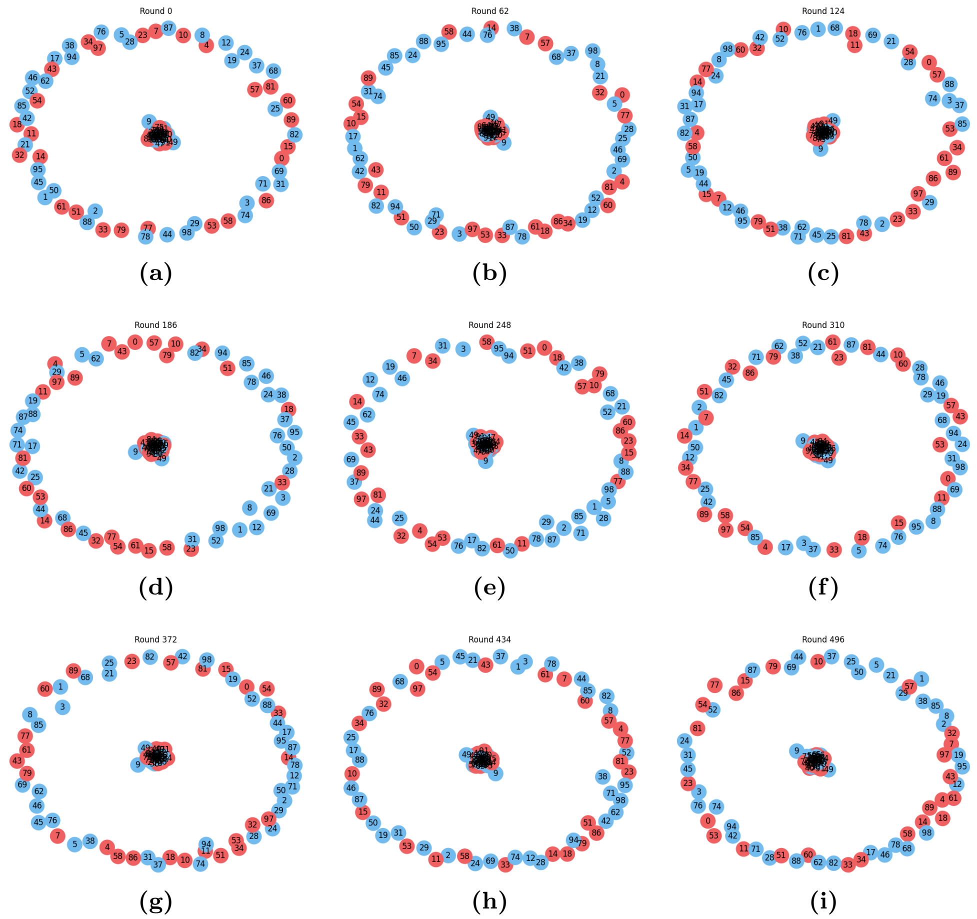


Figure 4.5: Evolution of cooperation in a multi-layer network across representative rounds. Each plot shows the network state at different rounds, with blue nodes representing cooperative entities and red nodes representing defecting entities. Node numbers represent the unique identifiers for each agent. We used an LLM consult interval of 33,000. The cooperative prevalence values indicate the percentage of cooperative agents within the network at each round. As the simulation progressed, we can observe varying levels of cooperation and the formation of clusters of cooperative and defecting agents, thus illustrating the dynamic nature of the multi-agent system. (**a**) cooperative prevalence = 57%; (**b**) cooperative prevalence = 49%; (**c**) cooperative prevalence = 49%; (**d**) cooperative prevalence = 45%; (**e**) cooperative prevalence = 60%; (**f**) cooperative prevalence = 58%; (**g**) cooperative prevalence = 57%; (**h**) cooperative prevalence = 53%; (**i**) cooperative prevalence = 55%.

We would like to emphasize that, while the EC framework has been demonstrated via a simplified simulation, we believe that the principles and mechanisms it encapsulates, such as adaptive learning, multi-layered interactions, and the use of LLM-based strategy recommendations, bear significant relevance to complex real-world scenarios. The ability of our framework to model and promote cooperation among diverse and adaptive agents provides a powerful tool to address various challenges in different contexts.

4.7 Implications for Business and Society

The EC framework, as demonstrated by our simulation, has far-reaching implications for both businesses and society as a whole. By promoting cooperation and fostering positive interactions between agents, the EC framework can be applied to a variety of real-world scenarios to optimize social welfare and enhance cooperation.

In the context of businesses, the EC framework can be used to model and improve cooperative behavior between employees, teams, or departments, potentially leading to increased productivity and efficiency within organizations. Moreover, the insights gained from the LLM-based strategy recommendations can inform decision-making processes and help organizations adapt to changing environments.

From a societal perspective, the EC framework can be applied to model and address pressing issues such as climate change, public health, and economic inequality. By encouraging cooperative behavior among individuals, communities, and nations, the EC framework can facilitate the development of sustainable solutions to these complex challenges.

Despite the promising results obtained from the EC framework, several limitations should be acknowledged. First, the simulation environment used in this study is a simplified representation of real-world systems. The assumptions made about agent behavior, network structure, and interactions may not fully capture the complexity of real-world situations. Additionally, the choice of LLMs and their implementation within the EC framework may also influence the outcomes observed in the simulation.

4.8 Conclusions

In this chapter, we have presented a comprehensive framework that integrates EC Theory, MASs, and LLMs to simulate and analyze the dynamics of cooperation and defection in complex environments. By incorporating heterogeneous agents, adaptive learning mechanisms, and LLM-based strategy recommendations, our framework provides a more realistic and flexible representation of HCI in MASs.

We have also discussed the implementation details of our simulation environment, including performance metrics, visualization techniques, and the use of intelligent sensors for data collection and real-time adaptation. Through the analysis of various simulation results, we have demonstrated the emergence of cooperation and defection patterns, the influence of LLM-based strategy recommendations, the robustness and resilience of the system under different conditions, and the utility of our visualization techniques for understanding multi-agent system dynamics.

Furthermore, we have discussed the broader implications of our findings for business and society, thereby highlighting the potential benefits and challenges associated with the integration of LLMs and MASs in various domains. We have also acknowledged the limitations of our current framework, including the incorporation of additional layers of interaction, more advanced LLM-based strategy formation mechanisms, and the development of more sophisticated visualization and analysis tools.

In our proposed framework, we extended the concept of HCI to encompass the interaction between human agents and AI-driven agents, such as LLMs, in complex Multi-Agent Systems. This extended interpretation of HCI aims to capture the intricate dynamics of cooperation and defection that arise when humans and AI collaborate, compete, or coexist in various domains. By integrating LLMs as a form of human–computer interface, we created a more adaptive and flexible representation of these interactions, where the LLM modifies the beliefs and strategies of human agents based on the information provided. This approach allows for a deeper understanding of the potential benefits and challenges associated with human–AI collaboration in complex environments and contributes to the development of more effective and efficient Human–Computer Interaction strategies in diverse real-world applications.

In conclusion, our study represents a significant step towards a deeper understanding of the interplay between humans and computers in cooperative and competitive settings. By integrating advanced AI technologies, such as LLMs, with well-established theories from game theory and MASs, we aim to pave the way for more effective and efficient Human–Computer Interaction and unlock the potential of intelligent agents to address a wide range of complex problems in various domains.

Future work should focus on refining the EC framework by incorporating more realistic models of agent behavior, interaction mechanisms, and network structures. This can be achieved through the integration of empirical data, as well as the application of advanced modeling techniques. Furthermore, the performance of different LLMs and their suitability for various contexts should be explored.

Additional areas of future work include the investigation of alternative learning processes, the development of more sophisticated visualization techniques, and the study of the EC framework's applicability to a broader range of real-world scenarios. By addressing these limitations and expanding upon the current work, the EC framework has the potential to significantly contribute to our understanding of cooperation and defection in MASs.

Bibliography

[1] Nash, J. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49. [CrossRef]

[2] von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 1944.

[3] Busoniu, L.; Babuska, R.; Schutter, B.D. A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2008**, *38*, 156–172. [CrossRef]

[4] Stone, P.; Kraus, S. To teach or not to teach? Decision making under uncertainty in ad hoc teams. In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, ON, Canada, 10–14 May 2010; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 2010; Volume 1, pp. 157–164.

[5] Shah, D.; Osiński, B.; Ichter, B.H.; Levine, S. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In Proceedings of The 6th Conference on Robot Learning, Auckland, New Zealand, 14–18 December 2022; Volume 205, pp. 492–504.

[6] Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Visual Language Maps for Robot Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023.

[7] Fehr, E.; Schmidt, K.M. Theories of fairness and reciprocity–evidence and economic applications. In *Advances in Economics and Econometrics*; Dewatripont, M., Hansen, L.P., Turnovsky, S.J., Eds.; Econometric Society Monographs; Cambridge University Press: Cambridge, UK, 2002; Volume 1, pp. 208–257.

[8] Granovetter, M. Economic action and social structure: The problem of embeddedness. *Am. J. Sociol.* **1985**, *91*, 481–510. [CrossRef]

[9] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

[10] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.

[11] Osborne, M.J. *An Introduction to Game Theory*; Oxford University Press: Oxford, UK, 2004.

[12] Myerson, R.B. *Game Theory: Analysis of Conflict*; Harvard University Press: Cambridge, MA, USA, 1991.

[13] Fudenberg, D.; Levine, D.K. *The Theory of Learning in Games*; MIT Press: Cambridge, MA, USA, 1998.

[14] Camerer, C.F. *Behavioral Game Theory: Experiments in Strategic Interaction*; Princeton University Press: Princeton, NJ, USA, 2003.

[15] Mitchell, M. *An Introduction to Genetic Algorithms*; MIT Press: Cambridge, MA, USA, 1996.

[16] Rosin, C.D.; Belew, R.K. New methods for competitive coevolution. *Evol. Comput.* **1997**, *5*, 1–29. [CrossRef] [PubMed]

[17] Axelrod, R. Advancing the art of simulation in the social sciences. In *Simulating Social Phenomena*; Conte, R., Hegselmann, R., Terna, P., Eds.; Lecture Notes in Economics and Mathematical Systems; Springer: Berlin/Heidelberg, Germany, 1997; Volume 456, pp. 21–40.

[18] Li, G.; Wang, G.; Wang, S. Two-population coevolutionary algorithm with dynamic learning strategy for many-objective optimization. *Mathematics* **2021**, *9*, 420. [CrossRef]

[19] Nowak, M.A.; May, R.M. Evolutionary games and spatial chaos. *Nature* **1992**, *359*, 826–829. [CrossRef]

[20] Gong, D.; Xu, B.; Zhang, Y.; Guo, Y.; Yang, S. A similarity-based cooperative co-evolutionary algorithm for dynamic interval multiobjective optimization problems. *IEEE Trans. Evol. Comput.* **2019**, *24*, 142–156. [CrossRef]

[21] Meselhi, M.A.; Elsayed, S.M.; Sarker, R.A.; Essam, D.L. Contribution based co-evolutionary algorithm for large-scale optimization problems. *IEEE Access* **2020**, *8*, 203369–203381. [CrossRef]

[22] Angeline, P.J.; Pollack, J.B. Evolutionary module acquisition. In Proceedings of the Second Annual Conference on Evolutionary Programming, La Jolla, CA, USA, 25–26 February 1993; pp. 154–163.

[23] Vakhnin, A.; Sopov, E. Investigation of improved cooperative coevolution for large-scale global optimization problems. *Algorithms* **2021**, *14*, 146. [CrossRef]

[24] Stone, P.; Veloso, M. Multiagent systems: A survey from a machine learning perspective. *Auton. Robot.* **2000**, *8*, 345–383. [CrossRef]

[25] Deng, Z.; Fan, J.; Shi, Y.; Shen, W. A coevolutionary algorithm for cooperative platoon formation of connected and automated vehicles. *IEEE Trans. Veh. Technol.* **2022**, *71*, 12461–12474. [CrossRef]

[26] Weiss, G. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*; MIT Press: Cambridge, MA, USA, 1999.

[27] Wooldridge, M. *An Introduction to MultiAgent Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2009.

[28] Ho, E.; Rajagopalan, A.; Skvortsov, A.; Arulampalam, S.; Piraveenan, M. Game Theory in defence applications: A review. *Sensors* **2022**, *22*, 1032. [CrossRef] [PubMed]

[29] Oroojlooy, A.; Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. *Appl. Intell.* **2022**, *53*, 13677–13722. [CrossRef]

[30] Shoham, Y.; Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*; Cambridge University Press: Cambridge, UK, 2008.

[31] Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624.

[32] Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2022**, arXiv:2104.13921.

[33] Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.

[34] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[35] Gupta, M.; Sharma, B.; Tripathi, A.; Singh, S.; Bhola, A.; Singh, R.; Dwivedi, A.D. n-Player Stochastic Duel Game Model with Applied Deep Learning and Its Modern Implications. *Sensors* **2022**, *22*, 2422. [CrossRef]

[36] Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

[37] Muglich, D.; Schroeder de Witt, C.; van der Pol, E.; Whiteson, S.; Foerster, J. Equivariant networks for zero-shot coordination. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6410–6423.

[38] Yang, M.S.; Schuurmans, D.; Abbeel, P.; Nachum, O. Chain of thought imitation with procedure cloning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36366–36381.

[39] Macy, M.W.; Willer, R. From factors to actors: Computational sociology and agent-based modeling. *Annu. Rev. Sociol.* **2002**, *28*, 143–166. [CrossRef]

Chapter 5

Socratic Video Understanding on Unmanned Aerial Vehicles

I. de Zarzà, J. de Curtò, Carlos T. Calafate. (2023). "Socratic Video Understanding on Unmanned Aerial Vehicles." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pendiente de asignación.

In this chapter, we propose a system for video understanding through zero-shot reading comprehension using Socratic Models. Specifically, we create a language-based world-state history of events and objects present in a scene captured by an Unmanned Aerial Vehicle (UAV). To achieve this, video footage from RYZE Tello microdrones is transmitted to a ground computer for further processing. The semantically rich information offered by Large Language Models (LLMs) enables open-ended reasoning, such as event forecasting with minimal human intervention, in a cost-effective robotic system. BLIP-2 is employed to answer a given set of instructional prompts, creating a log-state of objects, humans, and hazards that can be searched. Simultaneously, it suggests probable actions in the scene and can assist the human controller with an estimated best command.

5.1 Introduction

Unmanned Aerial Vehicles (UAVs), commonly known as drones, are employed in various applications, particularly those requiring mobility and versatility. The emergence of models that can equip these robotic platforms with real-time intelligence, either with or without the assistance of a GPU-enabled ground station or wired computer, has paved the way for open-ended reasoning in lightweight systems. This advancement also enables real-time event tracking and description, serving as a companion to human operators in potentially hazardous situations by suggesting estimated optimal commands for action.

Large Language Models (LLMs) [10, 1, 23] have emerged as valuable resources for addressing complex tasks across diverse disciplines. Their pre-trained capabilities give rise to numerous zero-shot applications, referring to models that do not require retraining and, consequently, eliminate the need for specific adaptation to particular scenarios. In this context, UAVs play a significant role in deploying real-world robots equipped with sensing and visual capabilities. Their interaction with humans is particularly interesting in a wide range of environments [2]. The aim of this chapter is to bridge the gap between state-of-the-art pre-trained models, which are based on vast data corpora, and affordable robots designed for human interaction. Our goal is to generate a comprehensive scene description in the form of a world-state history, specifically a log of all objects, humans, and likely actions per frame. Additionally, we seek to enable the ability to query LLMs regarding specific details of the ongoing events.

To achieve this, we employ a RYZE Tello drone, an affordable lightweight microdrone equipped with a high-definition camera. We configure this drone to transmit the captured video stream to a ground computer responsible for computation and world-state acquisition. Socratic Models (SMs) [24] represent a modular framework wherein tasks are formulated as language-based exchanges between multiple modules, and zero-shot capabilities are preserved through prompt engineering and guided multimodal interactions among the utilized models. In this work, we rely on BLIP-2 [13], a generic and efficient pre-training strategy that bootstraps vision-language pretraining from a frozen image encoder and frozen LLMs. BLIP-2 offers captioning capabilities, as well as generic image understanding in the form of Socratic interactions between models and users. Our objective is to develop an affordable robotic system that takes advantage of recent advances in Natural Language Processing (NLP) and

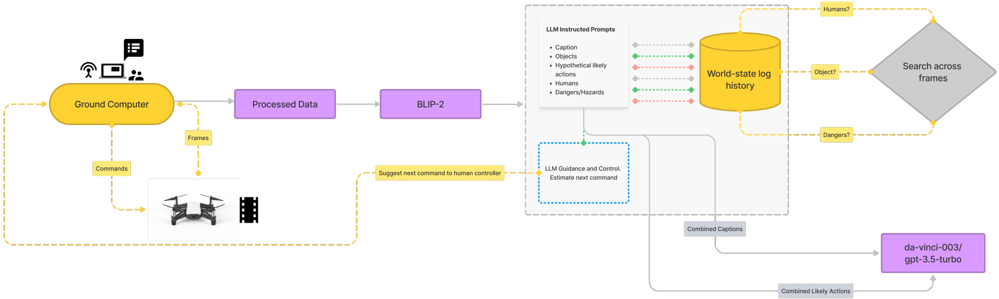


Figure 5.1: Diagram illustrating the pipeline for Socratic Video Understanding, which combines BLIP-2 and OpenAI da-vinci-003/gpt-3.5-turbo. The system can operate in two modes: either by collecting frames and conducting post-processing to generate a searchable world-state history or in real-time, providing an estimated best command suggestion for the human-in-the-loop.

Computer Vision while providing practical insights on the usability and best practices of these devices in real-world scenarios. For this purpose, we establish an experimental setup with the drone functioning as a flying camera that transmits video to the ground computer. A set of hard-coded guidelines in the form of open-ended question-answers is supplied to the system to be queried at each frame and then combined to generate a world-state log of activities, actions, and objects; refer to Figure 5.1 for a visual representation of the methodology.

The chapter is organized as follows: Section 5.2 provides an overview and state of the art of LLMs and VLMs. Next, in Section 5.3 we discuss the problem we are trying to characterize. Section 6.2 addresses the materials and methodology used, where we highlight the main contributions and the body of work of the publication. In Section 5.5 we evaluate the proposed methodology and design a set of experiments. Section 5.5.1 gives emphasis on the use of the pipeline for guidance and command. Finally, Section 6.4 presents the conclusions drawn after testing the proposed framework, discusses the limitations of actual LLM architectures, and suggests potential future work for further investigation.

5.2 Overview and related work

Large Language Models (LLMs) [7, 10, 1, 23] and Visual Language Models (VLMs) [18] are widely acknowledged as essential resources that enable intelligent systems to interact with humans in novel ways and perform complex tasks effortlessly. These models, also known as Foundation Models, have demonstrated remarkable performance across various tasks, including robotic manipulation [6, 25, 17], navigation, and guidance [24, 11]. They can also be integrated with other modules to address highly complex situations.

LLMs are a type of ML models trained on vast text datasets to generate human-like text. These models rely on Transformers [20, 8] to learn and adapt through training, aiming to discern statistical patterns and relationships in the data to produce plausible text. One of the most notable features of LLMs is their compelling ability to generate text that is virtually indistinguishable from text written by humans. Owing to their training on substantial volumes of text, these models can capture various linguistic patterns, such as grammar, syntax, and vocabulary, and create coherent and grammatically correct text ideal for Natural Language Processing (NLP) applications like translation, summarization, and text generation [5, 26].

VLMs, in contrast, are ML models trained on both text and image datasets to generate descriptive and informative text about visual content. The primary objective of VLMs is to learn the statistical patterns and relationships present in the training data and leverage this knowledge to generate natural language text describing the objects, people, and events depicted in an image or set of images. One of the significant strengths of VLMs is their ability to produce grounded text that is closely related to the visual content of an image. This means that the generated text provides detailed and informative descriptions of the visual elements, enabling the models to perform tasks such as image classification, object detection, and image captioning. VLMs have the potential to revolutionize image and video analysis, content generation, and NLP. In this chapter, we focus on Vision-Language Pre-training (VLP) [14, 12, 13], which emphasizes the use of these large pre-trained modules in a wide variety of tasks, and we combine them in a Socratic manner.

Furthermore, recent advancements have delved into the utilization of LLMs and VLMs in Socratic learning, wherein these models are integrated in a way that facilitates mutual learning, thereby augmenting their collective understanding of the environment. Such approaches have demonstrated potential across various applications, encompassing robotics and autonomous systems [9, 3]. The work presented in our chapter aligns with these cutting-edge developments and proposes an innovative pipeline that merges LLMs and VLMs using a Socratic approach to enable UAVs to attain semantic scene understanding.

5.3 Research problem

Advances in NLP and Computer Vision have enabled numerous use cases in which the analysis of visual cues equips agents with real-time intelligence in unparalleled ways. Socratic Models can simultaneously serve as object detectors, event forecasters, and anomaly or hazard detectors. In this context, tasks that were typically addressed using multiple backbones can now be characterized by a set of interconnected vision and language modules that communicate through prompts. Moreover, humans also play a role in this interaction by providing specific queries to the system, which can be either commands or actions to trigger a particular response, or questions about the specific situation captured by the sensors. Additionally, the system can assist the human in making appropriate decisions, such as determining the optimal command for the subsequent state. Drones are rapidly gaining popularity due to their potential applications in many real-world scenarios, including the deployment of AI techniques for semantic scene understanding harvesting state-of-the-art pre-trained models. This capability is crucial for applications such as robotics, surveillance, and autonomous driving, which necessitate the ability to analyze and interpret the meaning or significance of the objects, people, and events present in a scene.

The goal of this chapter is to confer the reader with a practical system implementation for UAV Socratic Video Understanding, utilizing state-of-the-art Large Language Models [19, 4, 18, 16]. We emphasize the widespread applicability of the techniques and the affordability of the hardware used. Our aim is to demonstrate the deployment of these instructable models in the context of drones, which can interact with the human on-the-loop to offer detailed descriptions, alerts, and guidance. Additionally, we are interested in highlighting the limitations of state-of-the-art models in accomplishing these tasks, as well as exploring the extent to which we can leverage interconnected large pretrained models and their actual understanding in real scenarios. As part of the proposed system for UAV semantic scene understanding, we have integrated BLIP-2 [13], a cutting-edge vision-language approach that builds on pre-trained unimodal models. To bridge the modality gap, they employ the Querying Transformer (Q-Former) pre-trained in two stages. The first stage involves learning vision-language representations using a frozen image encoder, followed by a vision-to-language generative learning stage using a frozen LLM. The Q-Former model architecture and two-stage pre-training process are integral components of our system, allowing us to efficiently process and analyze visual and textual data.

5.4 Materials and methodology

In our experiments, UAV footage from indoor and outdoor scenes has been used as the primary source of visual information for further processing, without relying on LiDAR, RADAR, or positioning information provided by an IMU. The captured videos were taken in uncontrolled environments, featuring a wide range of objects in the scenes, as well as spontaneous humans, traffic, and animals. Specifically, we consider three environments: an indoor setting in the form of a laboratory and corridor, and two outdoor scenes – a touristic avenue in front of a museum and an alleyway, as illustrated in Figure 5.2.

A language-based world-state history is composed by querying a Large Language Model. In particular we use BLIP-2 [13] and formulate a set of instructed prompts per frame, as follows:

a) Caption.

b) Which objects do you see in the image?

c) Which action is likely to happen in this image?

d) Are there humans in the scene?

e) Is the situation dangerous or are there some hazards?

The information provided by the Socratic Model as answer to these particular queries is stored and further processed for a post hoc analysis, visualization and interpretation.

Each of these instructed prompts is specifically chosen for a particular purpose. One of the goals is to obtain a comprehensive description of the scene by effectively combining the captions. A list of objects present in each frame pro-

Figure 5.2: The first and second rows depict outdoor scenes: the top row showcases a touristic avenue, while the latter features a university alleyway. The third and fourth rows correspond to indoor settings at the university, including a laboratory and a corridor, respectively. These diverse environments provide a wide range of scenarios for testing the proposed UAV Socratic Video Understanding system, allowing us to analyze its performance and capabilities in different real-world settings.

vides the user with a detailed log of the scene. The ability to forecast possible actions is especially valuable in many applications, where anticipating future likely events is essential for preventing danger or hazards. Additionally, the presence of humans serves as a reliable indicator of activity in a frame and can trigger further processing on specific frames. Similarly, the LLM's ability to assess whether a situation entails danger or hazards can be highly beneficial for initiating further event triggering in the form of a rapid response or an alarm. Given the LLM outputs for questions d) and e), regarding the presence of humans or dangers in a given frame, we train an SVM classifier to preprocess the text, transforming it into a feature-log containing only boolean indicators. Then, we plot the results over time for visualization, as shown in Figure 5.4. This information is particularly useful for analyzing the video content and triggering potential UAV reactions. It is important to note here that human and danger detection is pursued from the LLM perspective. Instead of using a traditional vision pipeline with an object detector, we leverage the LLM's ability to understand the scene and query its resources for further knowledge distillation. In the case of question b), the world-state log of objects can be utilized to search for the presence of specific instances of interest.

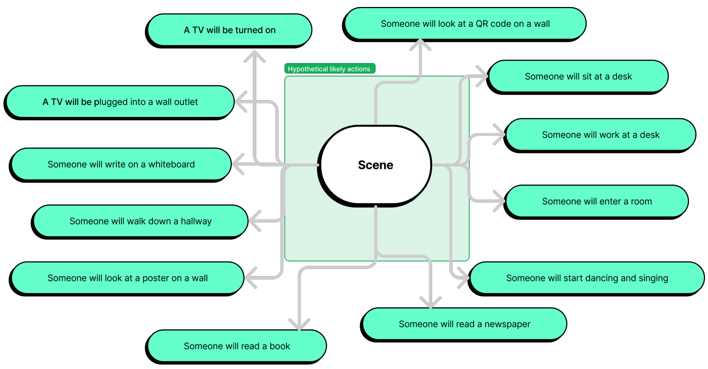


Figure 5.3: Conceptual plot of hypothetical likely actions given a video scene as suggested by BLIP-2. This visualization aims to provide an intuitive understanding of the potential actions that could occur within the scene based on the LLM's understanding of the situation. The plot demonstrates how the proposed UAV Socratic Video Understanding system can predict and anticipate events in the scene, providing valuable insights for various applications, such as safety monitoring, surveillance, and guidance.

For captioning, instructed prompt a), semantic scene understanding is achieved by combining the captions into a description using the API of OpenAI [19, 4], specifically the model da-vinci-003. We also employ the recently introduced API for ChatGPT [21], using the model gpt-3.5-turbo. Similarly, for likely actions (instructed prompt c), the results are combined into a hypothetical set of actions that could occur in the scene. Figure 5.3 provides a conceptual plot of this process. In this manner, we explore the interoperability of several frameworks in a Socratic way, where the outputs from BLIP-2 are combined and the API from OpenAI is called upon to complete the task.

5.5 Results

Experiments were conducted in both outdoor and indoor environments to evaluate the effectiveness of the proposed pipeline. Figure 5.4 displays the LLM indicators for human presence and potential dangers or hazards across four scenes, using a sampling rate of 1fps. A SVM is trained on the LLM outputs for danger and human presence (instructed prompts d) and e)) to generate boolean indicators, which are then plotted on a timeline for each frame. For example, the danger or hazard indicator is activated when the drone is in close proximity to the terrain or an object, while the human presence indicator is positive in every frame where a person appears.

Building on the same concept, instructed prompt b) is utilized to search through the frames for the presence of specific objects. Figure 5.5 presents the results for particular attributes, such as 'bike', 'tv', 'door', or 'bus', in both indoor and outdoor environments. As mentioned earlier, these techniques enable object detection capabilities from a NLP perspective, which underscores the potential for generalization across various applications of the proposed pipeline.

Instructed prompt (a) from all frames is combined to create a description using the API of OpenAI, GPT-3 model da-vinci-003 and gpt-3.5-turbo, providing textual information that accurately describes both indoor and outdoor scenes; see Tables 5.1 and 5.2. Readability metrics for LLM-enhanced text assessment, introduced in [7], are provided in Tables 5.3 and 5.6 to analyze the output of the improved gpt-3.5-turbo model, demonstrating very good generalization behavior.

The readability metrics [7] serve as indicators for the level of difficulty in understanding the text in the LLM-enhanced video descriptions for each environment. GUNNING Fog, Dale-Chall, ARI, Coleman-Liau, and Linsear Write have been employed as measures. GUNNING Fog measures the number of years of formal education required for a person to understand the text, with higher values indicating increased difficulty. Dale-Chall also provides a measure of text difficulty, with scores above 9 suggesting that the text may be challenging for some readers to comprehend. The ARI assesses text complexity, with higher scores signifying more difficult text. Coleman-Liau estimates text readability by measuring sentence length and average number of syllables per word. Lastly, Linsear Write estimates the years of formal education needed to understand the text. The results demonstrate that the LLM has a strong command of English, producing text that is rich yet not overly complex, enabling a broad audience to understand it.

The measures indicate that the model gpt-3.5-turbo has a very good grasp of English. For example, all scores Dale-Chall in Table 5.3 are higher than 9, suggesting that the texts are intended for an educated audience. Comparing these results to those reported in [7], where CLIP prefix for image captioning [18, 15, 16] and YOLOv7 [22] are combined with GPT-3, da-vinci-002, for UAV semantic scene understanding, it can be seen that the pipeline proposed using

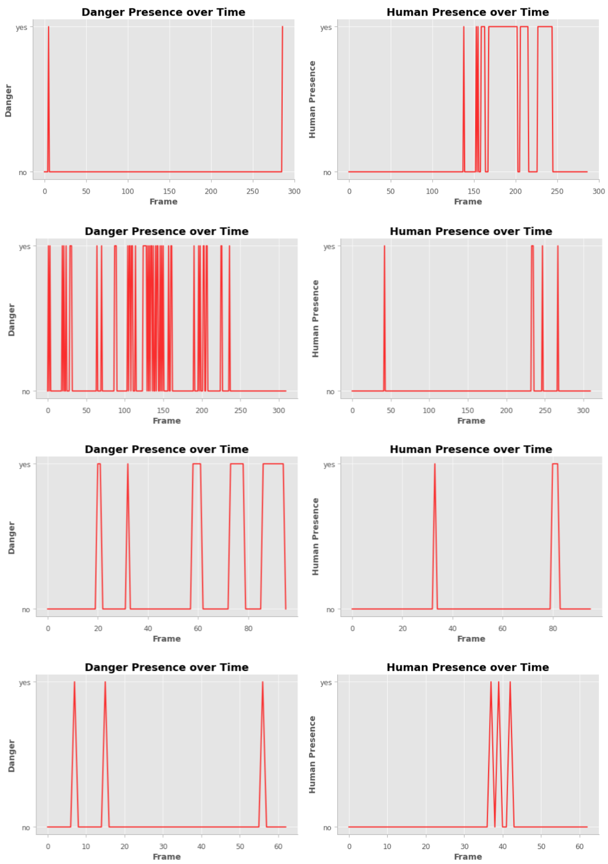


Figure 5.4: Indicators of presence of humans and dangers/hazards over time for the touristic avenue scene (first row), university alleyway (second row), laboratory (third row) and corridor (fourth row).

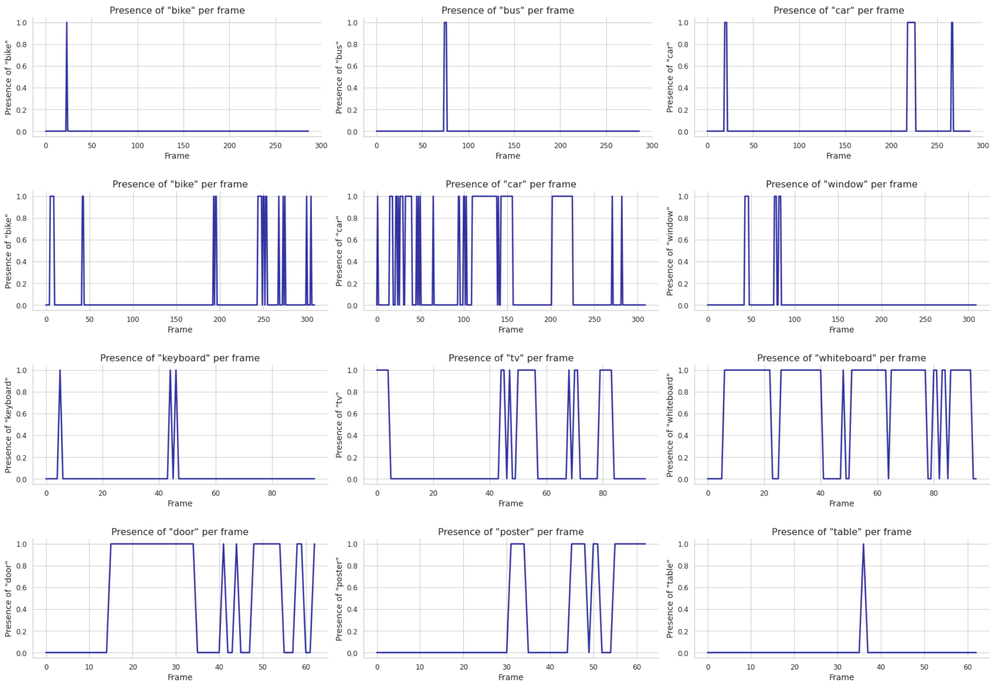


Figure 5.5: Presence indicators of specific objects using the world-state log provided by the LLM over time for the touristic avenue scene (first row), university alleyway (second row), laboratory (third row), and corridor (fourth row).

Table 5.1: Combined captions, instructed prompt a), for Videos [1] to [4] using the API of OpenAI, GPT-3 model: da-vinci-003.

Video	Description
[1]	A city park with a lot of grass and trees, a grassy area with a tree in the middle of it, a sign, a bench, a sculpture of the word love, a statue of a T-Rex and a giraffe, a man flying a kite, a person riding a skateboard down a city street, a van parked in front of an old building, and a lot of trash on the ground, all on a cloudy day.
[2]	A skateboarder is riding a skateboard down a sidewalk in front of a building, with a car parked in an empty parking lot next to a building with a metal fence and a fire hydrant, a bike rack, a lot of bicycles, and a bike tied to a pole, and a chain around it, in a city with a bus stop, a large tree in the middle of a parking lot, and a building with a clock tower in the background.
[3]	A small office with a whiteboard and a desk, a printer, two computer monitors on a desk, a whiteboard with writing on it, a box on a table, a TV mounted on the wall, and a trash can on the floor with a cat sitting next to it.
[4]	A long hallway in a building with white walls and posters on the wall, a TV mounted on the wall with wires attached to it, a whiteboard in a room with a trash can next to it, and a person sitting at a desk in an office.

BLIP-2 and GPT-3/ChatGPT yields more detailed text descriptions. Moreover, the highlighted text in Table 5.2 demonstrates that, for a target audience at the university level, we can expect to find nexus, composed sentences, subordinate clauses, and circumstantial components.

Instructed prompt (c) generates likely actions per frame, which are subsequently combined for event forecasting. The resulting text can be found in Tables 5.4 and 5.5, showcasing the exceptional forecasting capabilities of BLIP-2 when paired with either GPT-3 model da-vinci-003 or gpt-3.5-turbo. Notably, the newly introduced ChatGPT API, corresponding to the model gpt-3.5-turbo (Table 5.2), exhibits state-of-the-art performance for text summarization in terms of text quality, speed, and token capacity. The summary of likely actions provided by da-vinci-003 (Table 5.4) is structured and well-organized, while the output generated by gpt-3.5-turbo for this task (Table 5.5) is more descriptive and informative, albeit less concise.

Among the measures displayed in Tables 5.3 and 5.6, GUNNING Fog presents an ideal behavior, with all values falling within the range of $[7 - 12]$. This indicates that the generated texts are comparable in complexity to established publications in magazines and books, making them accessible to the general public. This holds true for both the descriptions of the combined captions and

Table 5.2: Combined captions, instructed prompt a), for Videos [1] to [4] using the API of OpenAI, GPT-3 model: gpt-3.5-turbo.

Video	Description
[1]	A bustling city park with green grass, towering trees, and unfortunately, a lot of trash scattered about. In the midst of it all, there's a popular grassy area where people walk and skateboard. A few interesting sights include a prominent sign and a striking sculpture. Oh, and don't forget the T-Rex statue that's causing quite a stir among park-goers. Meanwhile, a man flies a kite, a dog runs through the grass, and a person rides their skateboard down a cobblestone street. The park also features several benches and fountains, and there are plenty of places to take a break from the concrete jungle.
[2]	A collection of images showcasing streets, buildings, sidewalks, and empty parking lots in a city, including parked cars, bicycles, and skateboarders. Many of the buildings have doors and windows, and some have bushes, trees, grass, or fences in front of them. There are also several red and white poles, as well as a blue handicapped sign on the sidewalk.
[3]	A small office, computer room, or classroom with desks, chairs, whiteboards, a printer, and a mounted TV on the wall. Includes Samsung printers and related comparisons, with occasional presence of cats and trash cans.
[4]	A long hallway in various buildings with white walls and doors, mounted with numerous posters and whiteboards. Some rooms have desks and chairs, while others have TVs mounted on the wall with wires attached. Trash cans are seen next to the whiteboards in some rooms.

Table 5.3: Readability Metrics for Videos [1] to [4] for combined captions, instructed prompt a), gpt-3.5-turbo.

	Metric	**[1]**	**[2]**	**[3]**	**[4]**
Readability	**GUNNING Fog**	**8.82**	**11.95**	**11.51**	**7.78**
	Dale-Chall	9.4	9.97	12.37	10.7
	ARI	10.2	12.2	12.6	9.5
	Coleman-Liau	9.57	10.62	12.29	10.08
	Linsear Write	8.67	12.17	10.5	7.17

Table 5.4: Hypothetical likely actions, instructed prompt c), for Videos [1] to [4] using the API of OpenAI, da-vinci-003.

Video	Hypothetical likely actions (event forecast)
[1]	A person is walking on a sidewalk, walking in a park, throwing a frisbee, riding a skateboard, sitting on a bench, taking pictures of a Tyrannosaurus Rex statue, looking at a T-Rex statue, walking past a T-Rex statue, walking through the grass, and driving down the street, while a dog is playing with a frisbee, a group of people is walking in the park, and a car is parked in front of a building.
[2]	A person walks down the sidewalk, a car is parked in a parking lot, a car will drive into the parking lot, a car will drive down the street, a person will walk through the fence, a skateboarder will skate on the sidewalk, a skateboarder will skateboard down the street, a person is waiting for a bus, a person will ride a bike, a person will walk through the chained area, a person will try to enter the building, someone is going to use a broom to clean the sidewalk, someone is going to clean the sidewalk, a person enters a building and will use the wheelchair ramp, while a skateboarder is going to enter the same building, and a person will park their bike in the racks, sit on a bench, and put a tarp over the trash cans while a fire is extinguished.
[3]	A person will use a computer to watch a TV show, someone is working on a computer, writing on a whiteboard, working at a desk, typing on a computer, printing a document, working on a project, opening a box, using the printer, opening the door, walking into the room, and walking on the floor, while someone else is cleaning up the floor.
[4]	A TV will be turned on, plugged into a wall outlet, and connected to the wires; someone is writing on a whiteboard, walking down a hallway, looking at a QR code on a wall, sitting at a desk, working at a desk, and entering a room; a person is walking down the hallway, past a poster on a wall.

Table 5.5: Hypothetical likely actions, instructed prompt c), for Videos [1] to [4] using the API of OpenAI, gpt-3.5-turbo.

Video	Hypothetical likely actions (event forecast)
[1]	People enjoy various activities in the park including throwing a frisbee, walking their dog, riding a skateboard, sitting on benches, taking pictures of sculptures and statues, and walking on the grass and cobblestones. A group of people are also seen protesting while others are picking up trash. Cars drive down the street and are parked in front of buildings. The park also features statues of a tyrannosaurus rex and other sculptures, which attract visitors and tourists who take pictures and admire them.
[2]	People and vehicles moving around as usual in a busy urban environment with someone walking, a skateboarder riding and performing tricks, and cars parked in a lot or driving through the area. Some people are waiting for a bus or riding bikes, while others enter or exit buildings. A bike is parked in a planter or on a rack, and someone will use a wheelchair ramp.
[3]	In a busy room, people are multitasking - one person is using a computer to watch a TV show while another works on a project at the desk. Meanwhile, someone is writing on a whiteboard and someone else is about to print a document. A group of people collaboratively works on a project as students write on the whiteboard and use the printer. Amidst all the activity, someone will enter the room, open a box, and possibly clean up the floor while others walk on it.
[4]	A person will connect a TV by plugging it into a wall outlet, connect the wires, and turn it on. They will then work at a desk, sit down, and walk down the hallway passing posters on the wall with QR codes.

Table 5.6: Readability Metrics for Videos [1] to [4] for combined hypothetical likely actions, instructed prompt c), gpt-3.5-turbo.

	Metric	**[1]**	**[2]**	**[3]**	**[4]**
Readability	**GUNNING Fog**	**10.64**	**11.22**	**11.34**	**8.4**
	Dale-Chall	9.47	9.51	7.67	8.06
	ARI	13.4	10.9	10.2	7.6
	Coleman-Liau	12.3	8.36	7.83	4.7
	Linsear Write	11.5	12.33	12.5	10.5

the overall hypothetical likely actions.

The actual system operates offline, capturing a series of frames and processing them on a ground computer. Once the large pretrained model is loaded, answers per frame can be obtained within milliseconds, while processing thousands of frames may take several seconds. The system has the potential to function in real-time, with the main bottleneck being the frame capture and relay to the computer, which introduces a latency of a few milliseconds. In this context, the communication between the drone and the ground GPU plays a critical role in the performance of the system. Any delay in data transmission or processing due to network latency or limited bandwidth could affect real-time decision-making and responsiveness. To minimize this impact, it is crucial to optimize the communication protocol and data transfer rate. In Section 5.5.1, we propose using the LLM to assist in drone command by estimating the optimal action. In this case, minimizing latency is crucial, particularly when sending the image from the drone to the computer and then transmitting the text-based command from the computer to the drone. Strategies such as data compression, prioritization of critical information, or using edge computing devices on the UAV itself could be implemented to improve communication efficiency and help mitigate the latency issue and improve real-time performance. Although challenges remain, the latency could still be less than a second, making the system suitable for various real-time applications.

5.5.1 LLM guidance and control

The capability of prompting the LLM [9, 3] with queries could be further utilized for guidance and control. In this regard, it could be employed to suggest the next possible action to a human operating the UAV, or further integrated into the pipeline as a remote autonomous controller. For instance, using an instructed prompt such as: "Given this image taken by a drone, what's the best possible action: forward, backward, up, down, rotate left, or rotate right?" At each frame, BLIP-2 can be queried to determine the next best possible action based on the input frame. The drone then executes the action and processes another frame, and so on. In our system, we use this information to suggest the estimated best possible command to the human controller. Although the ability of BLIP-2 and state-of-the-art models to understand spatio-temporal data is still limited, and computations are performed per frame, the rapid advancements in the field ensure that this technology could soon empower autonomous robots and drones with intelligence in an unprecedented manner, achieving open-ended reasoning with visual cues. Thus, the system's capability to suggest the next best estimate command to the human controller is achieved through the interaction with the BLIP-2 model. At each frame, the system queries the BLIP-2 with an instructed prompt. The model processes the input frame and generates a response based on its understanding of the scene and the objects within it. This response represents the best possible action, as estimated by the model, which is then suggested to the human controller. The controller can consider the suggestion and decide whether to execute the recommended action or make a different decision. This process is repeated for each frame, allowing the human controller to continuously receive guidance from the model, enhancing the overall control and navigation of the UAV in real-time.

5.6 Conclusions and future work

Thorough experimentation has been conducted to assess UAV semantic scene understanding using an LLM pipeline based on BLIP-2 combined with OpenAI da-vinci-003/gpt-3.5-turbo. This methodology enables users to query the model about the environment, opening up a wide range of applications where Socratic video understanding can be achieved through human-in-the-loop interaction with a set of interconnected LLM modules that share information. Our system involves data collection from a RYZE Tello, which is then processed for knowledge acquisition using a series of instructed prompts. These prompts are stored and analyzed to perform further tasks by combining information from each frame into a global knowledge state that can be searched, for example, looking for particular objects, humans, presence of hazards, possible actions, and forecasting events across frames. The system is also capable of suggesting the next best estimated command to the human controller.

Potential future work could involve integrating positioning information into the pipeline, such as GPS coordinates or IMU data, along with data from internet-available maps, to enable event suggestions or site recognition. Additionally, the proposed pipeline is general and can be queried for a wide variety of information, allowing users to specify particular prompts tailored to their application of interest in a human-computer interaction context. Moreover, the system's capability to suggest an estimated best command based on raster images paves the way for LLM navigation and control directly within the pixel space. The limitations of the methodology stem from the LLM used to obtain answers to the instructed prompts. The capacity for open-ended reasoning depends on the model's ability to generalize. In this study, we observe that BLIP-2 can perform a wide variety of tasks in a zero-shot manner within real environments with changing conditions. However, newer models may be able to improve upon this and possess spatio-temporal knowledge of the scenes. In this regard, the LLMs under investigation can make reasonable guesses based on the instructed prompts, but they are not yet capable of making, for example, adequate probabilistic guesses, comprehending temporal aspects, or estimating length and size directly from visual data.

Bibliography

[1] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K., 2022. Flamingo: a visual language model for few-shot learning. arXiv:2204.14198 .

[2] Bonatti, R., Bucker, A., Scherer, S., Mukadam, M., Hodgins, J., 2021. Batteries, camera, action! learning a semantic control space for expressive robot cinematography. ICRA .

[3] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B., 2022. RT-1: Robotics transformer for real-world control at scale, in: arXiv preprint arXiv:2212.06817.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S.,

Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. Neural Information Processing Systems .

[5] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J., 2022. Scaling instruction-finetuned language models. arXiv:2210.11416 .

[6] Cui, Y., Niekum, S., Gupta, A., Kumar, V., Rajeswaran, A., 2022. Can foundation models perform zero-shot task specification for robot manipulation? L4DC .

[7] de Curtò, J., de Zarzà, I., Calafate, C.T., 2023. Semantic scene understanding with large language models on unmanned aerial vehicles. Drones 7. doi:10.3390/drones7020114.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR .

[9] Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P., 2023. PaLM-E: An embodied multimodal language model. arXiv:2303.03378 .

[10] Gu, X., Lin, T.Y., Kuo, W., Cui, Y., 2022. Open-vocabulary object detection via vision and language knowledge distillation. ICLR .

[11] Huang, W., Abbeel, P., Pathak, D., Mordatch, I., 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. ICLR .

[12] Li, D., Li, J., Le, H., Wang, G., Savarese, S., Hoi, S.C.H., 2022a. LAVIS: a library for language-vision intelligence. arXiv:2209.09019.

[13] Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv .

[14] Li, J., Li, D., Xiong, C., Hoi, S., 2022b. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: ICML.

[15] Li, X.L., Liang, P., 2021. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 .

[16] Mokady, R., Hertz, A., Bermano, A.H., 2021. ClipCap: CLIP prefix for image captioning. arXiv:2111.09734 .

[17] Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A., 2022. R3M: A universal visual representation for robot manipulation. RSS .

[18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. ICML .

[19] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. Technical report .

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

[21] Vemprala, S., Bonatti, R., Bucker, A., Kapoor, A., 2023. ChatGPT for Robotics: Design Principles and Model Abilities. Technical Report MSR-TR-2023-8. Microsoft. URL: `https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/`.

[22] Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696 .

[23] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W., 2022. Emergent abilities of large language models. arXiv:2206.07682 .

[24] Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., Lee, J., Vanhoucke, V., Florence, P., 2022. Socratic models: Composing zero-shot multimodal reasoning with language. arXiv:2204.00598 .

[25] Zeng, A., Florence, P., Tompson, J., Welker, S., Chien, J., Attarian, M., Armstrong, T., Krasin, I., Duong, D., Wahid, A., Sindhwani, V., Lee, J., 2020. Transporter networks: Rearranging the visual world for robotic manipulation. CoRL .

[26] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L., 2022. OPT: Open pre-trained transformer language models. arXiv:2205.01068 .

Chapter 6

Area Estimation of Forest Fires using TabNet with Transformers

I. de Zarzà, J. de Curtò, Carlos T. Calafate. (2023). "Area Estimation of Forest Fires using TabNet with Transformers." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pendiente de asignación.

In this chapter, we propose a novel approach for estimating the burned area of forest fires using the TabNet transformer-based architecture. Forest fires pose a significant threat to ecosystems, and accurate estimation of the affected area is essential for effective disaster management and resource allocation. We conducted a comprehensive analysis of various Machine Learning (ML) and Deep Learning (DL) methods, including Random Forest, Neural Networks, Neural Architecture Search (NAS), TabNet with Transformers, and Self-Supervised Learning with Autoencoders, to identify the most accurate and efficient model for area estimation. Our experiments employed a publicly available dataset, UCI Forest Fires, containing a combination of meteorological, geospatial, and categorical data. We implemented a thorough preprocessing pipeline that included handling categorical variables, standardization, and feature engineering.

6.1 Introduction

Forest fires [15, 10] have increasingly become a global concern, causing significant damage to ecosystems, wildlife, human life, and property. Accurate and timely estimation of the burned area in forest fires is crucial for effective disaster management, resource allocation, and planning of mitigation strategies. Recent advancements in Machine Learning (ML) and Deep Learning (DL) [8] have opened up new possibilities for addressing this critical issue. This chapter presents a comprehensive study of various methodologies for predicting the burned area of forest fires, with a particular focus on the TabNet transformer-based model [1, 14, 7]. In the past, traditional statistical and ML methods, such as linear regression and decision trees, have been employed for forest fire area estimation [9]. However, the complex interactions between various meteorological, geospatial, and categorical variables make these methods less effective in capturing the underlying patterns in the data. As a result, researchers have been exploring more sophisticated DL techniques that can better model the intricate relationships among these variables.

In this chapter, we have conducted a thorough study of diverse ML and DL approaches, including Random Forest [2], Neural Networks [3, 5], Neural Architecture Search (NAS) [17, 12, 13], Transformers [14, 7, 16, 4], and Self-Supervised Learning [6] with Autoencoders. We utilized a publicly available dataset containing a rich set of meteorological, geospatial, and categorical attributes. Our preprocessing pipeline involved handling categorical variables, standardization, and feature engineering to ensure the optimal input format for the various models. We employ grid search and hyperparameter optimization strategies for the Random Forest, and DL models. Furthermore, we delve into the application of NAS, an advanced technique for discovering NN architectures tailored to the specific problem of forest fire area estimation.

In this study, we introduce a novel approach to forest fire area estimation by leveraging the power of the transformer-based TabNet model, which has rarely been applied in this domain. Our work innovatively combines meteorological, geospatial, and categorical data in a cohesive model that captures the complex, interrelated factors contributing to fire spread. The novelty of our work lies not only in the application of TabNet transformers to this particular problem but also in the comprehensive, methodical approach we have adopted. We believe our methodology, which includes a thorough statistical analysis, data preprocessing and transformation, and rigorous model training and evaluation

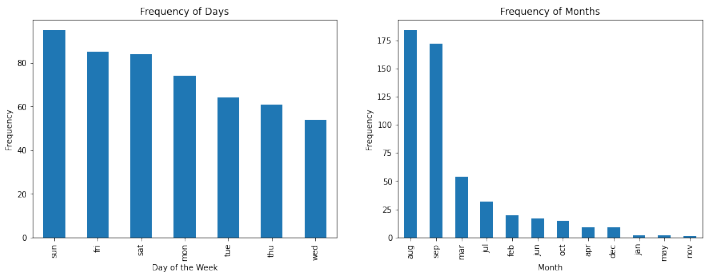


Figure 6.1: Basic statistical analysis: categorical variables. Bar graphs.

procedures, provides a unique and valuable reference point in the field of forest fire prediction.

This chapter is organized as follows: Section 6.2 provides a detailed description of the dataset, its features, and the preprocessing steps undertaken, as well as a thorough statistical analysis of the data, feature engineering and dimensionality reduction. Section 6.3 discusses the application of TabNet [1] and other methodologies under study. Finally, Section 6.4 concludes the study and provides directions for future research in the field of forest fire area estimation using ML and DL techniques.

6.2 Materials and methodology

We first prepare a basic statistical analysis, where we calculate the frequency and create a bar graph for categorical variables, as shown in Figure 6.1. In the case of numeric variables, we calculate basic descriptive statistics such as mean, median, and standard deviation. We also illustrate the histograms for all variables involved, as presented in Table 6.1 and Figure 6.2. Finally, for geospatial variables, we display a two-dimensional histogram, as depicted in Figure 6.3.

Categorical variables: bar graphs for the "day" and "month" attributes show the frequency distribution of wildfires between days of the week and months of the year. From the graphs, we observe that fires occur more frequently on weekends (Saturday and Sunday), possibly due to increased human activities or recreational visits to the forest during these days. The months with the highest frequency of fires are August and September, which can be attributed

Table 6.1: Descriptive Statistics for Numeric Variables.

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
count	517.0	517.0	517.0	517.0	517.0	517.0	517.0	517.0	517.0	517.0	517.0
mean	4.7	4.3	90.6	110.9	547.9	9.0	18.9	44.3	4.0	0.0	12.8
std	2.3	1.2	5.5	64.0	248.1	4.6	5.8	16.3	1.8	0.3	63.7
min	1.0	2.0	18.7	1.1	7.9	0.0	2.2	15.0	0.4	0.0	0.0
25%	3.0	4.0	90.2	68.6	437.7	6.5	15.5	33.0	2.7	0.0	0.0
50%	4.0	4.0	91.6	108.3	664.2	8.4	19.3	42.0	4.0	0.0	0.5
75%	7.0	5.0	92.9	142.4	713.9	10.8	22.8	53.0	4.9	0.0	6.6
max	9.0	9.0	96.2	291.3	860.6	56.1	33.3	100.0	9.4	6.4	1090.8

to the summer season and its higher temperatures, lower humidity, and dry conditions, making the forest more susceptible to fires.

Numerical variables: histograms and descriptive statistics of numerical variables, as shown in Figure 6.2 and Table 6.1, give us insight into their distributions and central tendencies:

- The “FFMC" (Fine Fuel Moisture Code) variable is slightly skewed to the left, with the majority of values concentrated around 90–95, indicating that fine fuel moisture conditions are generally high, which makes the forest more prone to fires.
- The “DMC" (Duff Moisture Code) variable has a right-skewed distribution with a long tail, meaning that there is a wide range of moisture content in the organic layers.
- The “DC" (Drought Code) variable is also skewed to the right, showing that drought conditions vary across the dataset, with some areas experiencing high levels of drought.
- The “ISI" (Initial Spread Index) variable shows a right-skewed distribution with a long tail, indicating that the rate of fire spread varies significantly, with some cases having a high rate of spread.
- The variable “temp" (temperature) shows an almost normal distribution, but slightly skewed to the right, with most temperatures between 15 and 25 degrees Celsius.
- The “RH" (relative humidity) variable is skewed to the left, indicating that lower humidity levels are more common in the dataset.

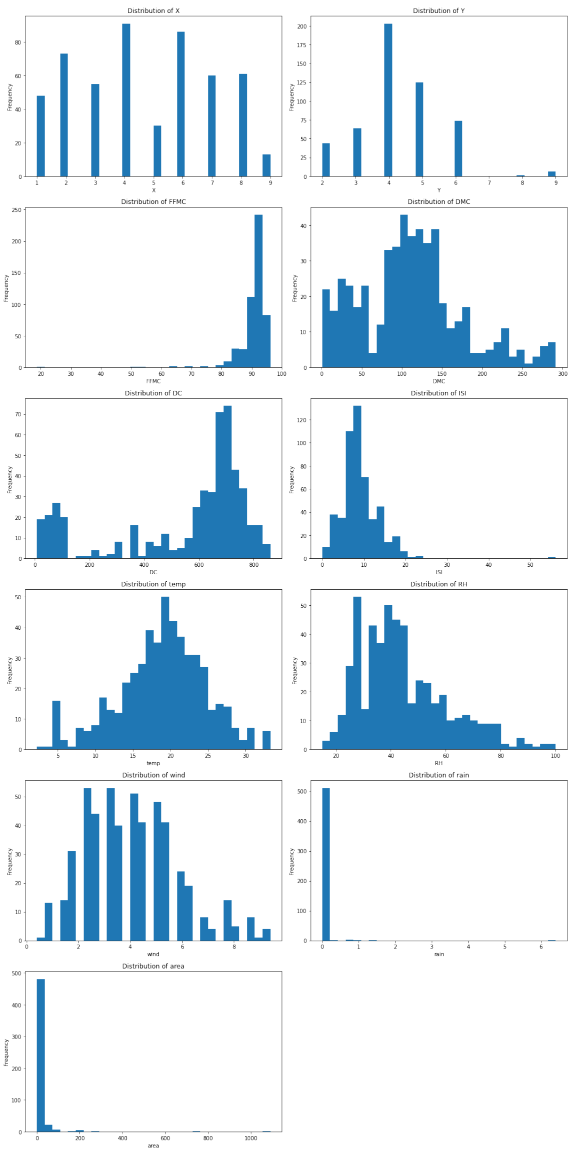


Figure 6.2: Basic statistical analysis: numeric variables. Plot of the histograms for all variables involved.

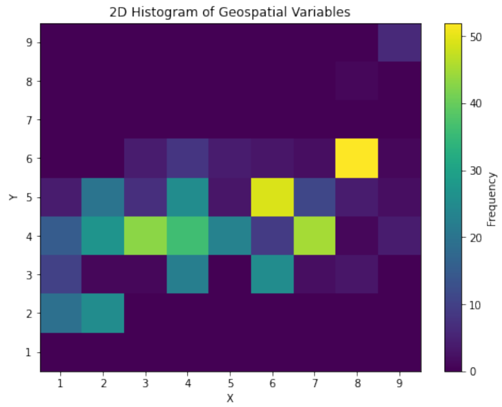


Figure 6.3: Basic statistical analysis: geospatial variables. Plot of the two-dimensional histogram.

- The “wind" variable is slightly skewed to the right, with most wind speeds between 2 and 6 km/h.
- The “rain" variable is heavily skewed to the right with a long tail, showing that most cases have little or no rain, but there are some cases with a significant amount of rain.

Geospatial variables: the 2D histogram of the geospatial variables “X" and “Y" provides information on the spatial distribution of wildfires, as depicted in Figure 6.3. There is a greater concentration of fires in the central region of the park (around coordinates $X = 4$, $Y = 4$) and the southeastern region (around coordinates $X = 7, Y = 4$), which indicates that these areas could be more prone to fires. There are also some areas with less fire incidence, such as the northwest and northeast corners of the park, which could be due to different types of vegetation, topography, or other factors that affect fire susceptibility.

From the superimposed histograms, as shown in Figure 6.4, it is not immediately apparent that the categorical attributes “day” and “month” have a strong relationship with the target variable “area”. However, some observations can be made: for the “day" variable, the histograms for different days appear to have similar shapes and distributions. Although the frequency of fires varies slightly between days, there does not appear to be a clear difference in the distribution of “area" values for each day of the week. For the variable “month", some differences can be observed. August and September, which have the highest

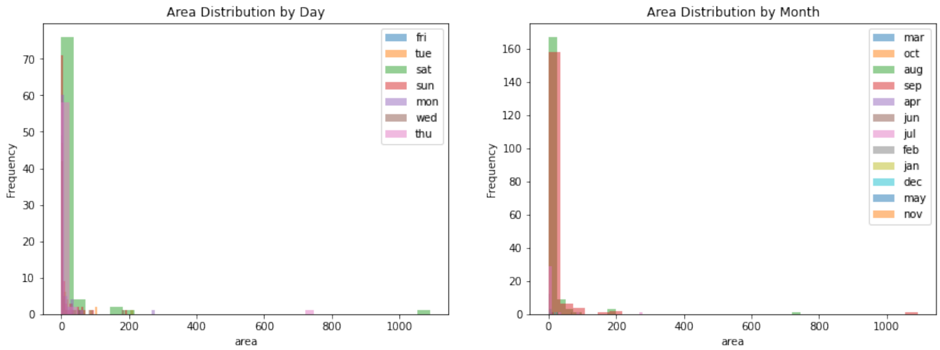


Figure 6.4: Exploratory Analysis. A graph for each categorical variable where there is overlaid for each category a histogram of the values of the target variable.

fire frequency, show a greater concentration of smaller "area" values. However, there are no clear patterns indicating that a specific month consistently leads to larger or smaller burned areas. Given these observations, the categorical attributes "day" and "month" may not have a significant impact in predicting the target variable "area". However, it is important to note that visual inspection alone may not be sufficient to definitively determine the relationship between these attributes and the target variable.

To better assess the importance of these categorical attributes, additional feature selection and statistical analysis techniques can be used. For example, we can also use correlation measures such as Cramér's V or perform a one-way ANOVA test to assess whether there are significant differences in mean "area" values between different categories of categorical variables. In addition, ML algorithms can be applied to evaluate the importance of these attributes during the modeling process. For each category of the categorical variables we compute the mean and standard deviation values of the target variable, see Tables 6.2 and 6.3. We compute then the pairwise correlation between all numerical attributes, including geospatial variables, see Figure 6.5. The resulting correlation matrix shows the linear relationships between all pairs of numerical variables, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value of 0 indicates no linear correlation. In Figure 6.6 we graph the relationships between all pairs of numeric variables in the dataset (including geospatial variables).

Table 6.2: Mean and Standard Deviation of "area" for Each Day.

Day	Mean	Standard Deviation
Fri	5.26	9.95
Tue	12.62	33.30
Sat	25.53	121.97
Sun	10.10	25.94
Mon	9.55	33.48
Wed	10.71	30.00
Thu	16.35	94.57

Table 6.3: Mean and Standard Deviation of "area" for Each Month.

Month	Mean	Standard Deviation
Mar	4.36	9.06
Oct	6.64	13.23
Aug	12.49	60.20
Sep	17.94	87.39
Apr	8.89	18.79
Jun	5.84	16.38
Jul	14.37	50.05
Feb	6.28	12.03
Jan	0.00	0.00
Dec	13.33	6.23
May	19.24	19.24
Nov	0.00	0.00

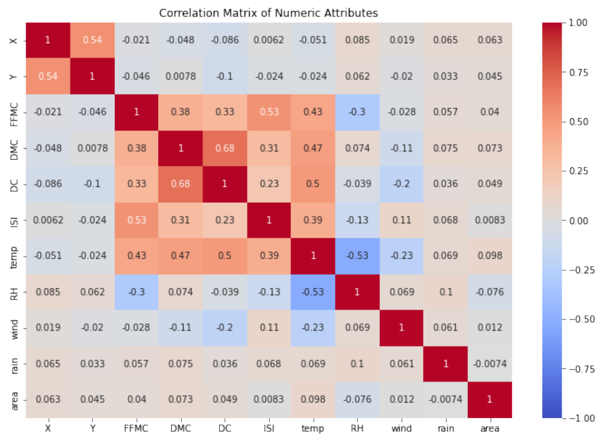


Figure 6.5: Pairwise correlation coefficients between all numerical attributes.

We then identify the 2 attributes that have the strongest correlation with the response, and the 3 that have the weakest correlation (higher or lower correlation coefficient in absolute value), as can be seen in Table 6.4. To observe and analyze the correlations graphically, we present, for each of the 5 identified attributes, a scatter plot with the attribute on the x-axis and the response on the y-axis. Also, we add a linear regression plot to each graph that fits the points, as shown in Figure 6.7.

Looking at the scatterplots, we can see the correlations that we have identified numerically to some extent, as can be seen in Figure 6.7. For the strongest correlations: the first attribute with the strongest correlation, although not

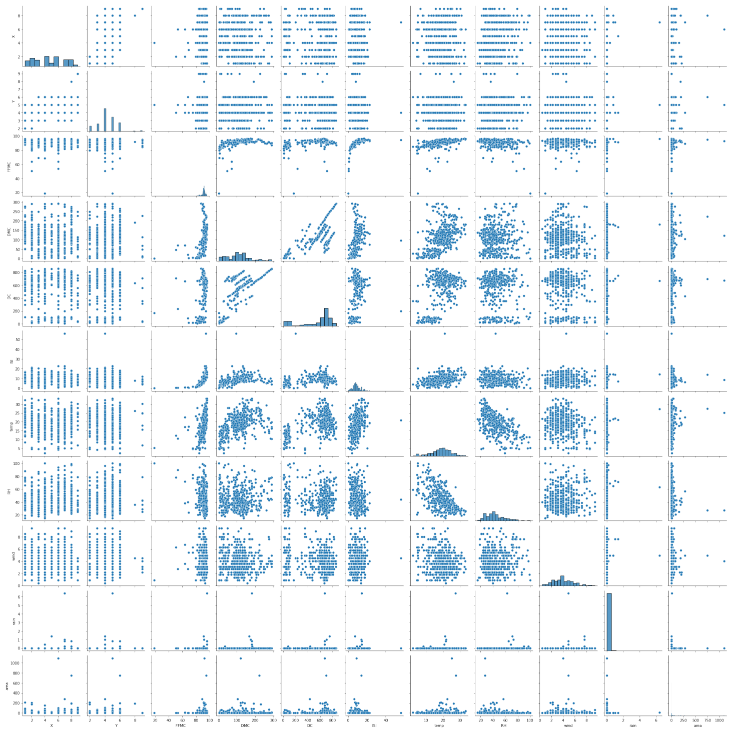


Figure 6.6: Graph with the relationships between all pairs of numeric variables in the dataset.

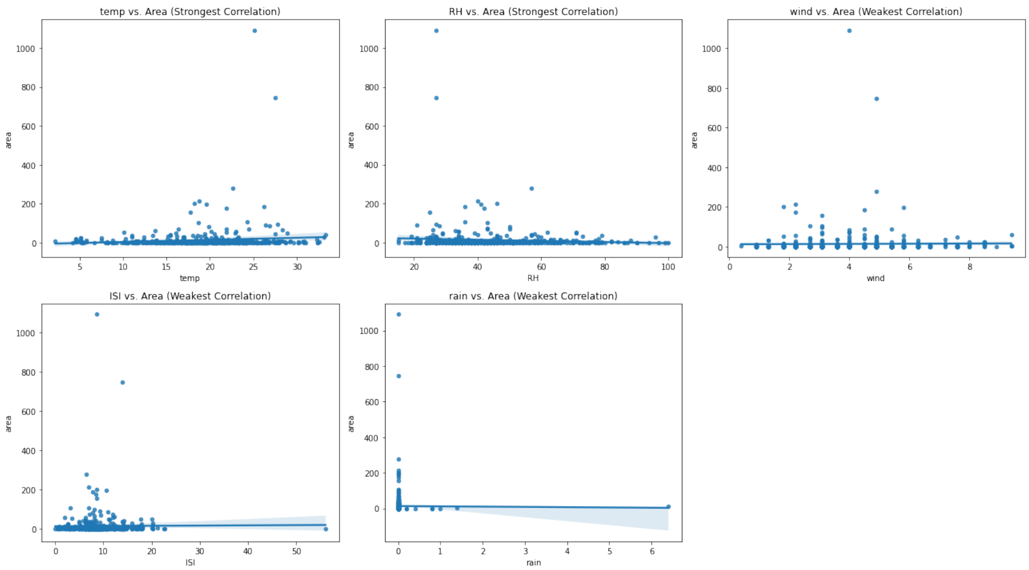


Figure 6.7: Scatter plot with the selected attribute on the x-axis and the response on the y-axis. We add a linear regression plot to each graph that fits the points.

Table 6.4: Strongest and Weakest Correlations with “area".

Variable	Correlation	Category
Temp	0.097844	Strongest
RH	-0.075519	Strongest
Wind	0.012317	Weakest
ISI	0.008258	Weakest
Rain	-0.007366	Weakest

very strong, shows a somewhat discernible pattern in the scatterplot. As the value of the attribute increases, we can see a slight increase in the “area" values, indicating a positive correlation between the two variables. The second attribute with the strongest correlation, however, does not show a very clear pattern in the scatterplot. The data points are scattered and it is difficult to identify a strong relationship with the “area" variable. For weaker correlations: in the first scatterplot of the weaker correlations, the data points are widely scattered with no apparent trend, confirming the weak correlation between

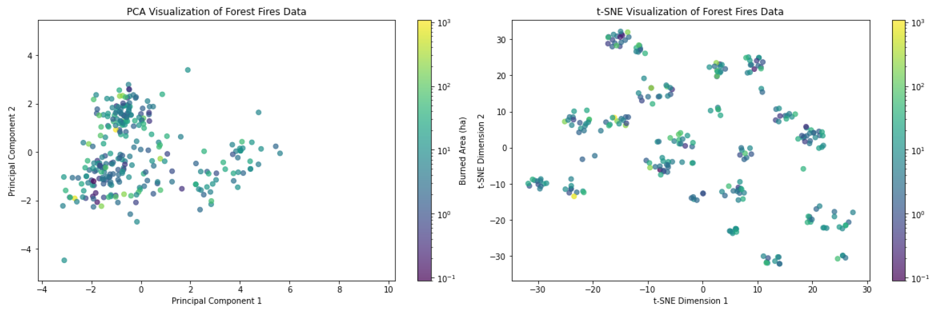


Figure 6.8: Projection to two attributes of the original descriptive attributes using PCA (left) and t-SNE (right).

the attribute and the “area" variable. The second scatterplot for the weaker correlations also shows a similar pattern, with no clear trend or relationship between the attribute and the “area" variable. The third scatterplot for the weakest correlations shows a similar pattern to the other weak correlations, with scattered data points and no clear trend or relationship to the “area" variable.

It is important to note that correlation only measures linear relationships, and there may be non-linear relationships between variables that are not captured by correlation coefficients. Additionally, the presence of potential outliers can also affect correlation values and scatterplots.

For the purpose of visually checking the distribution of the target variable considering all the descriptive attributes at once, we reduce the dimensionality of the problem to only two attributes which will be the projection of the original descriptive attributes using Principal Component Analysis (PCA) and t-SNE [11], as can be observed in Figure 6.8. Looking at the two graphs, it is not very clear whether the dimensionality reduction has worked well in both cases. Both the PCA plot and the t-SNE show no distinct clusters or patterns that would allow us to say with confidence that the new dimensions explain the variation in the target variable (area). The distribution of color (burned area) appears to be somewhat random in both plots, indicating that using only two novel dimensions might not be sufficient to differentiate between large and small fire areas.

6.3 Model comparison and results

In this section, we present the various models and techniques applied to the UCI dataset forest fires, along with their respective results. Our approach starts with basic statistical analysis and progressively moves to more advanced ML and DL methods. The primary goal is to understand the potential of each method in estimating the burned area and compare their performance.

The performance of each model has been evaluated based on their Mean Squared Error (MSE) on the training and test sets. Among all the models, TabNet [1] has emerged as the best-performing model, achieving the lowest MSE on the test set, as shown in Table 6.5 and Figure 6.9. This result indicates that TabNet's architecture is well-suited for the forest fire area estimation problem, as it efficiently captures the complex relationships between the variables in the dataset.

We begin with a thorough statistical analysis of the dataset, including descriptive statistics, as shown in Table 6.1, histograms, as depicted in Figures 6.1, 6.2, 6.3, and 6.4, and correlation analysis, as can be seen in Figure 6.5, 6.6 and 6.7 and Table 6.4. The categorical and numerical attributes are explored to understand their distributions and relationships with the target variable, area. This analysis provides valuable insights into the dataset and helps inform the choice of appropriate ML models. The dataset contains two categorical variables, 'month' and 'day.' We transform these categorical variables into binary variables using the "get_dummies" function in the pandas library. This transformation results in a dataset with additional binary columns representing each category. We standardize all descriptive attributes to have a mean of 0 and a standard deviation of 1 using the "StandardScaler" from the sklearn preprocessing module. This step ensures that all features have the same scale, improving the performance of the subsequent ML algorithms. We separate the dataset into training (70%) and test (30%) to evaluate the performance of our models. The train-test split is done after preprocessing to avoid data leakage, ensuring that the test set remains unseen and not influenced by any transformations applied to the training set.

Random Forest regression was selected as our first ML model due to its robustness and ease of implementation. We preprocessed the dataset by transforming categorical variables into binary variables and standardizing numerical attributes. The model was trained and evaluated on both training and test subsets, and the results showed good performance on the training data but overfitting to the training data, leading to a lower performance on the test subset. We use grid search over the parameters: n_estimators, max_depth, min_samples_split, min_samples_leaf, and bootstrap.

A DL model using feedforward neural networks was implemented to capture complex patterns in the dataset. Several architectures were explored, and the model was trained and evaluated on the same training and test subsets used for Random Forest. The results indicated that the DL model showed better generalization compared to the Random Forest. We use grid search over the parameters: epochs, batch_size, optimizer, hidden_layers, and neurons.

To further improve the model performance, we employed NAS to automatically discover good NN architectures for our specific problem. NAS searches for the best architecture by optimizing the model's structure and hyperparameters using techniques such as reinforcement learning and evolutionary algorithms. The discovered architecture was then trained and evaluated on the same training and test subsets used for the Random Forest and DL models. The results demonstrated that the NAS model outperformed Random Forest but not TabNet. We use AutoKeras with 100 epochs per configuration.

We also investigated the potential of self-supervised learning using an autoencoder, a type of neural network that can learn efficient data representations by encoding and decoding the input data. We trained the autoencoder on our dataset and used the learned feature representations as input to a subsequent regression model. While this approach showed promise, the performance was comparable to the NAS model but inferior to TabNet, although it showcased the potential of using unsupervised learning techniques for feature extraction and representation in regression tasks.

We employed the model TabNet [1], a DL architecture specifically designed for tabular data, which leverages the power of attention mechanisms and feature selection. TabNet has shown excellent performance in a variety of tasks, including regression problems like our current area estimation of forest fires. The model was implemented using the PyTorch library and optimized through hyperparameter tuning.

Here is a brief analysis of the scores obtained in Table 6.5 and Figure 6.9 (the lower the MSE, the better the model's performance):

Random Forest has the lowest training MSE of 1616, indicating that it performed the best during training. However, it has a relatively high testing MSE of 8044, which suggests the model might be overfitting the training data and not generalizing well to new, unseen data. DL model has a training MSE of 2184 and a testing MSE of 7870. Although the training performance is not as good as Random Forest, the testing performance is better, suggesting that this model generalizes better to new data. The NAS model has a training MSE of 2267 and a testing MSE of 7924. Its performance is slightly worse than the DL model, but it still generalizes reasonably well to new data. In the case of Self-supervised Learning with Autoencoders, the model has a training MSE of 2318 and a testing MSE of 7989. The performance is somewhat similar to the NAS model, but the testing error is slightly higher, indicating that it might not be the best choice for this problem. TabNet has a training MSE of 2319 and the lowest testing MSE of 7781 among all the models. This suggests that TabNet generalizes the best to new, unseen data in the context of forest fire area estimation; which is particularly interesting as it is a transformer-based architecture adapted to work with tabular data, a domain where traditional techniques in many cases outperform other methodologies.

The results of this study can be used to inform the development of more robust and efficient models for estimating the burned area of forest fires and for understanding the environmental impact of these events.

Table 6.5: MSE Train and Test results for the methodologies under study.

Method	MSE Train	MSE Test
Random Forest	1616	8044
Deep Learning	2184	7870
NAS	2267	7924
Self-supervised Learning with Autoencoders	2318	7989
TabNet	2319	7781

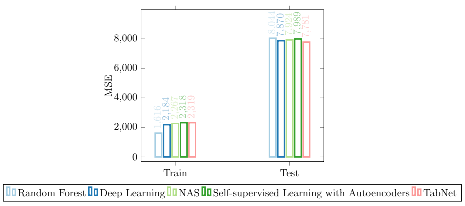


Figure 6.9: Comparison in bar plot of MSE Train and Test results.

6.4 Conclusions and future work

In this chapter, we presented a comprehensive study of several ML and DL models applied to the problem of estimating the burned area of forest fires. Our investigation began with a thorough statistical analysis of the dataset, which informed our subsequent choices of models and techniques. Among the models explored, TabNet demonstrated the best performance, significantly outperforming other approaches such as NAS, Self-Supervised Learning with Autoencoders, DL and Random Forests.

Our results highlight the potential of TabNet in solving complex regression problems and its superiority in obtaining very good estimates for the specific problem at hand. However, there are several avenues for future work that could further improve the performance and applicability of our models:

- Incorporating additional data sources: integrating external data, such as weather data or satellite imagery, could provide a richer context for the models and potentially improve their performance in estimating the burned area.
- Investigating alternative NAS techniques: while our NAS implementation delivered promising results, exploring other NAS methods, such as differentiable architecture search or Bayesian optimization, could lead to even better model architectures.
- Exploring other self-supervised learning techniques: although the autoencoder-based approach did not outperform the NAS model, investigating other self-supervised learning techniques could yield valuable insights and contribute to better feature extraction and representation in the problem domain.
- Investigating ensemble methods: combining the predictions from multiple models, such as NAS, transformers, and autoencoders, could lead to better overall performance by leveraging the strengths of each individual model and mitigating their weaknesses. Ensemble techniques, such as stacking or bagging, may be particularly beneficial in improving the robustness and generalization of our predictions.
- Assessing model performance on a wider range of datasets: to better understand the generalizability of our models, it would be valuable to test their performance on other forest fire datasets from different geographi-

cal regions and time periods. This would allow us to identify potential limitations and refine our models accordingly.

- Further exploration of transformer-based models: adapting other transformer-based models to the forest fire area estimation problem could potentially yield even better results, given their recent success in various fields.

In conclusion, our study demonstrates the potential of advanced ML and DL techniques, particularly a transformer-based architecture, for estimating the burned area of forest fires. We believe that further research in this area can lead to more accurate and robust models, ultimately contributing to better forest fire management and prevention efforts.

Bibliography

[1] Arik, S.O., Pfister, T., 2021. TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence 35, 6679–6687. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16826, doi:10.1609/aaai.v35i8.16826.

[2] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. p. 785–794. URL: https://doi.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

[3] Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., Yang, S., 2017. AdaNet: Adaptive structural learning of artificial neural networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR. pp. 874–883.

[4] de Curtò, J., de Zarzà, I., Calafate, C.T., 2023. Semantic scene understanding with large language models on unmanned aerial vehicles. Drones 7. doi:10.3390/drones7020114.

[5] de Zarzà, I., de Curtò, J., Calafate, C.T., 2022. Detection of glaucoma using three-stage training with efficientnet. Intelligent Systems with Applications 16, 200140. URL: https://www.sciencedirect.com/science/article/pii/S2667305322000771, doi:https://doi.org/10.1016/j.iswa.2022.200140.

[6] Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding, in: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics. pp. 4171–4186.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR .

[8] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. CVPR .

[9] Kalantar, B., Ueda, N., Idrees, M.O., Janizadeh, S., Ahmadi, K., Shabani, F., 2020. Forest fire susceptibility prediction based on machine learning models with resampling algorithms on remote sensing data. Remote Sensing 12. URL: `https://www.mdpi.com/2072-4292/12/22/3682`, doi:`10.3390/rs12223682`.

[10] Kurbanov, E., Vorobev, O., Lezhnin, S., Sha, J., Wang, J., Li, X., Cole, J., Dergunov, D., Wang, Y., 2022. Remote sensing of forest burnt area, burn severity, and post-fire recovery: A review. Remote Sensing 14. URL: `https://www.mdpi.com/2072-4292/14/19/4714`, doi:`10.3390/rs14194714`.

[11] van der Maaten, L., Hinton, G.E., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605.

[12] Pham, H., Guan, M., Zoph, B., Le, Q., Dean, J., 2018. Efficient neural architecture search via parameters sharing. ICML .

[13] Tan, M., Chen, B., Pang, R., Vasudevan, V., Le, Q.V., 2019. Mnasnet: Platform-aware neural architecture search for mobile. CVPR .

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

[15] Xie, Y., Peng, M., 2019. Forest fire forecasting using ensemble learning approaches. Neural Computing and Applications 31. doi:`10.1007/s00521-018-3515-0`.

[16] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L., 2022. OPT: Open pre-trained transformer language models. arXiv:2205.01068 .

[17] Zoph, B., Le, Q., 2017. Neural architecture search with reinforcement learning. ICLR .

Chapter 7

Detection of glaucoma using three-stage training with EfficientNet

I. de Zarzà, J. de Curtò and Carlos T. Calafate. (2022). "Detection of glaucoma using three-stage training with EfficientNet." Intelligent Systems with Applications, vol(16), 200140. DOI: 10.1016/j.iswa.2022.200140

This chapter sets forth a methodology that is based on three-stage-training of a state-of-the-art network architecture previously trained on Imagenet, and iteratively finetuned in three steps; freezing first all layers, then re-training a specific number of them and finally training all the architecture from scratch, to achieve a system with high accuracy and reliability. To determine the performance of our technique a dataset consisting of 17.070 *color cropped samples of fundus images, and that includes two classes, normal and abnormal, is used. Extensive evaluations using baselines models (VGG16, InceptionV3 and Resnet50) are carried out, in addition to thorough experimentation with the proposed pipeline using variants of EfficientNet and EfficientNetV2.*

7.1 Introduction

Intelligent systems based on data-driven techniques have been proposed in recent years to solve a wide variety of tasks with unprecedented level of success. In the case of medical data, applications where methods that rely on computer vision and neural networks are used have proven to be very effective; [22, 16, 4, 2], showing performance levels that are similar or even better than human assessment.

In this thesis, we propose a three-stage training mechanism to design a reliable system to detect and assess glaucoma; [6, 5], an irreversible neuro-generative eye disease that, according to the World Health Organization (WHO), affects more than 65 million people around the globe. Early detection and treatment are of utmost importance to prevent loss of visual capacity.

The system introduced achieves state-of-the-art performance on the application under consideration, and the methodology is general enough to be used in other clinical cases or widespread vision applications where learning highly descriptive features from raw pixel intensities is crucial. The design principles take into consideration performance, reliability, statistical significance, platform-aware latency and FLOPS needed to accomplish the task; with the ultimate goal to propose an expert system that could be seamlessly integrated with the clinical equipment (e.g. retinograph) for early diagnostic and treatment.

Our network achieves a mean average percentage F1-score across folds of 96.6 using EfficientNet-B0 (with standard deviation of 3.7) and EfficientNet-B4 (with standard deviation of 2.0), where the best F1 on a given fold is 99 on B0 and 98 on B4. For the case of EfficientNetV2, V2-B3 achieves a mean average F1-score across folds of 95.7 (with standard deviation of 2.3) and V2-S of 95.4 (with standard deviation of 1.6), where the best F1 on a given fold is 98 for both V2-B3 and V2-S. These results significantly outperform the baselines; VGG16 (83.2), InceptionV3 (91.1) and ResNet50 (88.9), and are also clearly better than the state-of-the-art reported results found in the literature, [6].

Code and data used throughout the chapter is released publicly under the badge initiative on reproducibility by Code Ocean[1]. A detailed notebook ad-

[1] Permanent link to reproducible capsule: doi.org/10.24433/CO.8342269.v1

dressing all the stages of the methodology, as well as the dataset used, can be found in a runnable capsule environment.

The remainder of this chapter is organized as follows: in the next section we present an overview and state of the art; then an exhaustive description of the data and the methods is provided, together with visualization, preliminary study, evaluation and selection of the best model with thorough experimentation. Finally, conclusions and further work are discussed.

7.2 Overview and state of the art

Several approaches to address retinal imaging problems are introduced, showing both traditional techniques based on hand-crafted features and also CNN based methodologies. A brief discussion of each procedure is provided as well as the type of data used in the experimentation. We broaden the analysis by mentioning the state-of-the-art techniques used in our work, that unlike previous publications take carefully into account accuracy and number of parameters.

In [3] they propose a semi-supervised learning framework based on bag of words for early detection of glaucoma. In [7] they assess the pathology by the use of random tree classification, although the experiments are only reported on a dataset of 45 samples. [32] details the use of PCA and BAYES classifier. [18] proposes the use of CNN to predict bounding boxes with their corresponding class probability and confidence score, where initialization is done using k-means clustering. [30] goes beyond this approach and applies a pathology-aware feature visualization approach for the diagnostic, where the method relies heavily on Generative Adversarial Networks (GANs). [8] uses UNet++ in [35] to segment the Optic Disc and Optic Cup using feature extraction at several fields of view and then a gradient boosting decision tree to do the screening of glaucoma. Traditional methods have also shown to be effective to tackle related medical imaging problems, [13, 33, 34]. For a meticulous analysis of several approaches see [1], where both a description of handcrafted methods and techniques based on deep convolutional networks [31, 21, 17] are presented. However, although most methods provide a well-thought effective methodology to address the problem, the majority of them have a shortfall on the used data, as are tested on very small datasets with limited statistical significance.

The architecture proposed is built on EfficientNet in [27] and EfficientNetV2 in [28], using a three-stage training mechanism that broadens the finetuning steps proposed in [6]. These architectures are built using Neural Architecture Search (NAS), [36, 37, 20], in particular EfficientNet uses the AutoML MNAS framework presented in [26] that optimizes the networks for accuracy and efficiency (FLOPS) and is based on previous work in [23] and [26], but with a larger base model. Evaluation of the family of models is done from B0 to B5 in the case of EfficientNet, and from B0 to B3, S and M in EfficientNetV2. We use transfer learning from ImageNet to the particular application under study and see that the models achieve high accuracy with a reduced number of training parameters, compared to other state-of-the-art methodologies. Concomitant approaches in retinal image classification show the adequacy of the family of models EfficientNet for the given task, [29, 19, 15, 14, 10, 9].

7.3 Data and methods

The dataset under consideration consists on 17.070 fundus images, which are digitalized photographs of the posterior part of the eye, with positive (abnormal) and negative (normal) samples of the pathology. The data is divided into 10 folds of 1.707 instances, each one with its corresponding sets of training, validation and testing. The sets are relatively balanced to reduce the number of false negatives. These samples are obtained using retinography, and thus their characteristics in terms of illumination and intensity are very particular and relatively homogeneous among all instances; such aspect is central for the correct detection of the samples. For this reason, using the raw pixels without normalization confers the network with significantly better generalization than when using min-max normalization, or normalization with standard deviation, as these types of preprocessing cause loss of information. This observation is very important as any type of non-linear transformation that affects or alters the brightness of the samples can severely degrade the performance of such a system; this can hold also when dealing with other medical data where image intensity is crucial.

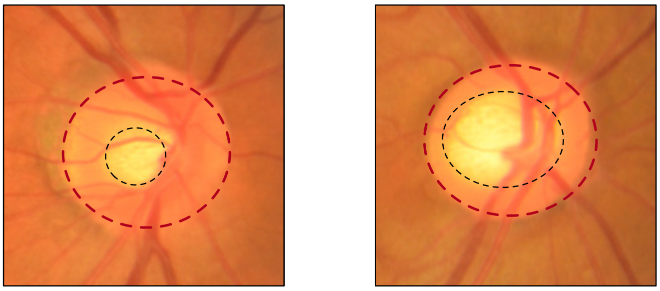


Figure 7.1: Example of negative (normal) and positive (abnormal) samples. Highlighted inner circular region corresponds to Optic Cup and outer circular region to Optic Disc. Samples that are glaucomatous (right; with severe pathology) present abnormal size of the Optic Cup respect to normal samples (left). [Source: two random samples from the dataset under study.]

7.3.1 Fundus images

The samples are cropped to improve the sensitivity of the detector. The disease is characterized by an abnormal size of the Optic Cup, with respect to the Optic Disc, see Figure 7.1. This is the reason why many earlier approaches were based on the Cup/Disc Ratio (CDR). As our approach is data driven, there is no need to use handcrafted intermediate features as feature selection. A random subset of the data is shown in Figure 7.2, as well as, detailed statistics in Table 7.1 and Figure 7.3.

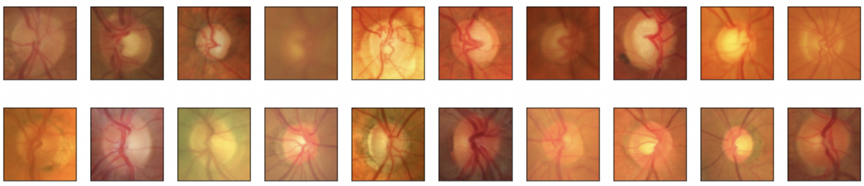


Figure 7.2: Visual exploration of the samples of the pathology glaucoma; first row: positive, second row: negative.

Table 7.1: Statistics of the dataset consisting on 17.070 fundus images with positive (P: abnormal) and negative (N: normal) samples. The data is distributed into 10 folds (0 to 9) of 1.707 samples each with corresponding train, validation and test.

		Fold										
		0	1	2	3	4	5	6	7	8	9	**Total**
Train	(N)	754	740	739	743	746	758	754	642	748	733	7452
	(P)	625	639	640	636	633	621	625	737	631	646	6338
Validation	(N)	83	88	83	85	81	71	84	82	80	82	819
	(P)	71	66	71	69	73	83	70	72	74	72	721
Test	(N)	82	91	97	91	92	90	81	100	91	104	919
	(P)	92	83	77	83	82	84	93	74	83	70	821
Total		1707	1707	1707	1707	1707	1707	1707	1707	1707	1707	17070

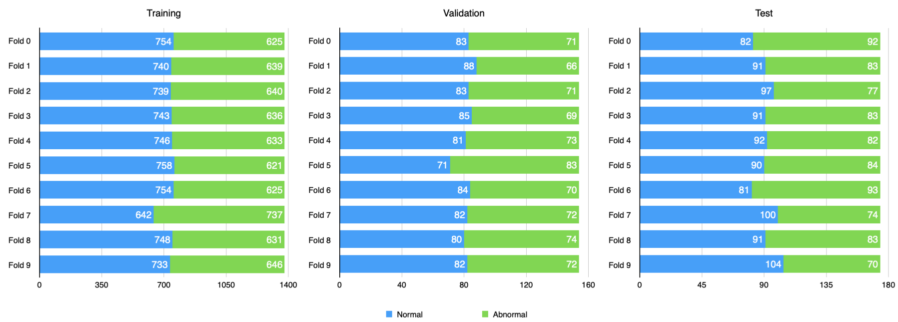


Figure 7.3: Statistics of the dataset using a bar plot for sets of training, validation and testing.

7.3.2 Preliminary study and methodology

The preliminary study focuses on Fold 0 to set forth a design methodology that will serve as guiding principle of the chapter.

The methodology under study proposes a three-stage training system (see Figure 7.4 for a visual description) that consists on the following procedure in only one Fold of the data:

1. Start from a model trained on Imagenet, and only re-train the last added layers (GlobalAveragePooling2D, BatchNorm, Dropout and Fully Connected) of the system.

2. Use the weights from the previous stage to initialize a model that unfreezes a number of layers of the previous model (excluding BatchNorm), and retrain the system.

3. Use the weights from the previous iteration and retrain the whole network. Evaluate the classification report based on both the F1-score and confusion matrix to select the best hyperparameters.

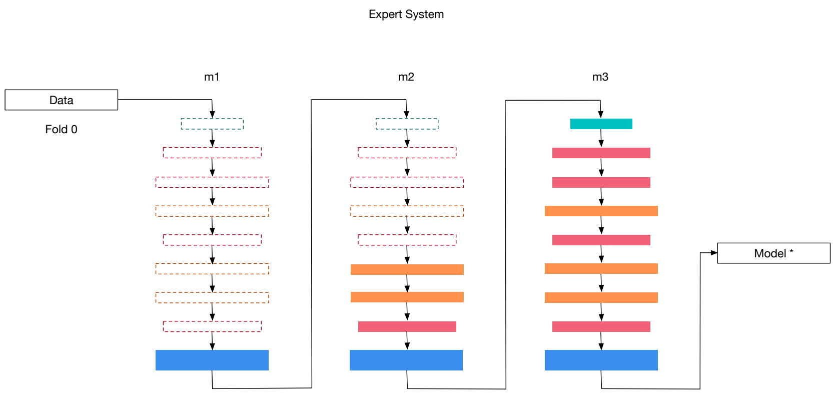


Figure 7.4: Visual description of the proposed three-stage training system in one fold of the data. Transfer learning from ImageNet is put in place. Color layers are re-trained. In particular, weights from m1 are used to initialize the network when training m2, which unfreezes a given number of layers from the full model (in our application 20, keeping layers BatchNorm untrained). Afterwards, weights from m2 are used to initialize the network when training m3, which retrains the whole architecture. Finally, in the evaluation stage, the weights obtained (Model*) are then fed into 10-fold crossvalidation to retrain the network for each fold, and select the best model according to F1-score. The procedure is robust against hyperparameter choices.

Experimentation is based on baseline models (VGG16 in [24], InceptionV3 in [25] and ResNet50 in [11]) and then extended to variants of EfficientNet; [27], and EfficientNetV2; [28].

EfficientNet-B0 base model consists on the following layers, see Table 7.2. In this particular example, model m1 freezes all layers from stages 1 to 8, and retrains only layers corresponding to stage 9. Model m2 starts from the learned weights on m1 and retrains a subset of layers going backwards, in our case 20, while keeping BatchNorm layers untrained. Finally model m3 starts from the weights of m2 and retrains the whole network. The network proceeds in the same way with the case of variants of EfficientNetV2 (see Table 7.3 for a description of the architecture) and model baselines (VGG16, InceptionV3 and ResNet50).

The family of models EfficientNet and EfficientNetV2 is a compositional stack of modules MB and Fused-MB Convolutions (denoted MBConvn and Fused-MBConvn). These modules consist on the following inner operators:

- **MBConvn**: a 1×1 convolution, followed by a depthwise 3×3 convolution, a SE module in [12], and finally another 1×1 convolution.
- **Fused-MBConvn**: a 3×3 convolution, followed by a SE module and finally a 1×1 convolution.

Table 7.2: EfficientNet-B0, baseline network. Each row describes a stage c with $\hat{L}_c$ layers, with input resolution $\langle \hat{H}_c, \hat{W}_c \rangle$ and output channels $\hat{C}_c$.

Stage c	**Operator** $\hat{\mathcal{F}}_c$	**Resolution** $\hat{H}_c \times \hat{W}_c$	**# Channels** $\hat{C}_c$	**# Layers** $\hat{L}_c$
1	Convn3x3	224×224	32	1
2	MBConvn1, k3x3	112×112	16	1
3	MBConvn6, k3x3	112×112	24	2
4	MBConvn6, k5x5	56×56	40	2
5	MBConvn6, k3x3	28×28	80	3
6	MBConvn6, k5x5	14×14	112	3
7	MBConvn6, k5x5	14×14	192	4
8	MBConvn6, k3x3	7×7	320	1
9	Convn1x1 & Pooling & FC	7×7	1280	1

Table 7.3: EfficientNetV2-S, example architecture. Extension to EfficientNet using both MB and Fused-MB Convolutions. Each row describes a stage c with $\hat{L}_c$ layers, with given stride and output channels $\hat{C}_c$.

Stage c	**Operator** $\hat{\mathcal{F}}_c$	**Stride**	**# Channels** $\hat{C}_c$	**# Layers** $\hat{L}_c$
0	Convn3x3	2	24	1
1	Fused-MBConvn1, k3x3	1	24	2
2	Fused-MBConvn4, k3x3	2	48	4
3	Fused-MBConvn4, k3x3	2	64	4
4	MBConvn4, k3x3, SE0.25	2	128	6
5	MBConvn6, k3x3, SE0.25	1	160	9
6	MBConvn6, k3x3, SE0.25	2	256	15
7	Convn1x1 & Pooling & FC	-	1280	1

Worthy of mention is the fact that EfficientNet-B0 achieves state-of-the-art performance while keeping the number of parameters to train bounded to the same levels as ResNet50.

Tables 7.4, 7.5 and 7.6 show the accuracy in validation and training for the given three-stage training mechanism: m1 corresponding to the first stage where only the last layers are trained, m2 to the second stage where a number of layers are unfrozen, and m3 to the third stage where the whole network is retrained.

Confusion matrices of the corresponding models are shown in Tables 7.7, 7.8 and 7.9, where we can see that there is a clear performance increase due to the three-stage procedure, causing the number of false negatives to be drastically reduced. Although for this particular task we consider F1-score as the comparison metric, confusion matrices allow for the computation of other measures such as error-rate, accuracy, specificity, sensitivity, and precision.

Numerical progression of F1-score can be observed in Tables 7.10, 7.11 and 7.12, where the number of trainable and non-trainable parameters are reported for each method level under study.

Table 7.4: Three-stage training system for several model baselines. Accuracy in Fold 0.

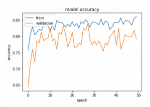
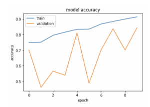
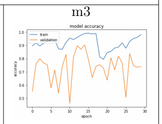
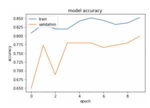
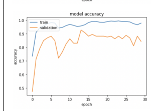
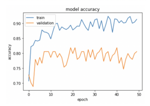
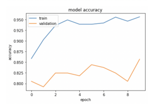
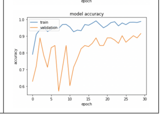
		Method level		
		m1	m2	m3
Model	VGG16			
	InceptionV3			
	ResNet50			

Table 7.5: Three-stage training system for several variants of EfficientNet. Accuracy in Fold 0.

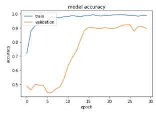
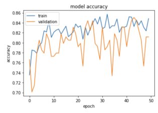
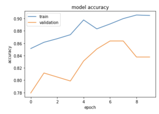
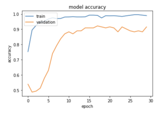
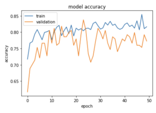
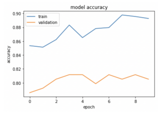
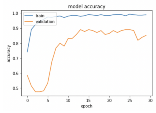
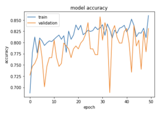
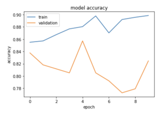
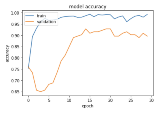
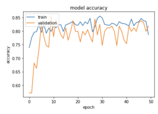
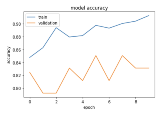
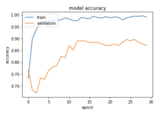
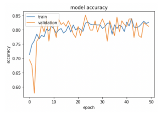
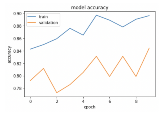
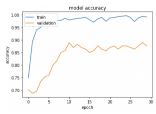
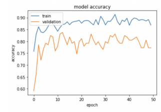
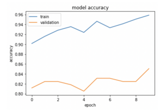
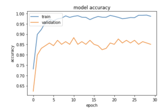
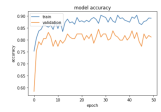
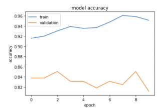
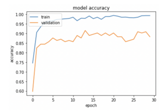
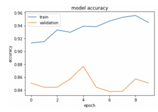
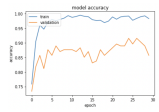
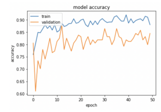
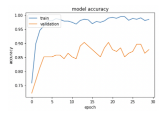
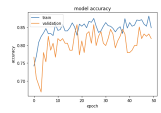
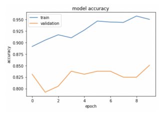
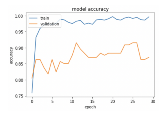
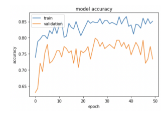
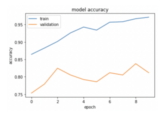
		Method level		
		m1	m2	m3
EfficientNet	B0			
	B1			
	B2			
	B3			
	B4			
	B5			

Table 7.6: Three-stage training system for several variants of EfficientNetV2. Accuracy in Fold 0.

		Method level		
		m1	m2	m3
EfficientNetV2	B0			
	B1			
	B2			
	B3			
	S			
	M			

Regarding the confusion matrices for the three-stage training system, we can observe that the expected behavior in terms of incorrect cases is having a higher number of false positives than false negatives. This is appropriate for designing a system to detect glaucoma, as the principle is to be able to always detect the disease if it is present, as the pathology is irreversible and early treatment can considerably improve the condition of the subject.

Table 7.7: Three-stage training system. Confusion Matrix Baseline Models in Fold 0. VGG16, InceptionV3 and ResNet50.

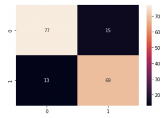
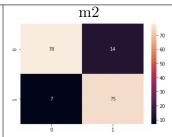
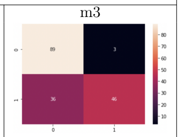
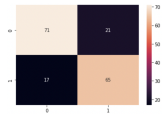
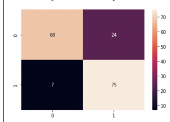
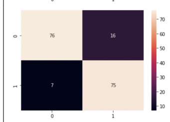
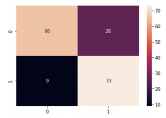
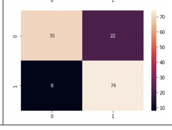
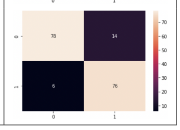
		Method level		
		m1	m2	m3
Model	VGG16			
	InceptionV3			
	ResNet50			

Table 7.8: Three-stage training system. Confusion Matrix EfficientNet in Fold 0.

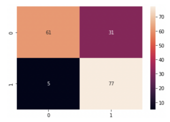
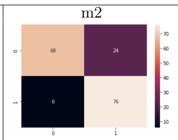
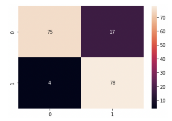
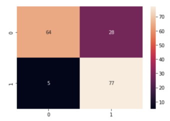
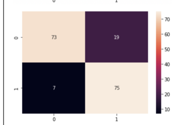
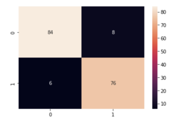
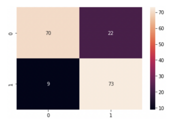
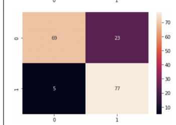
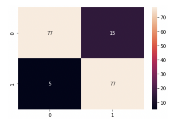
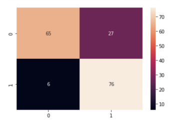
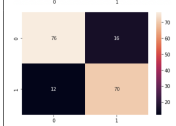
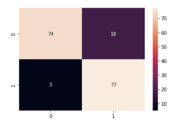
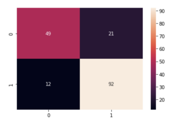
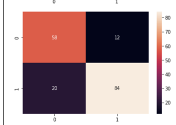
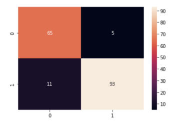
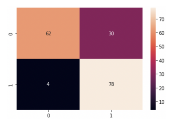
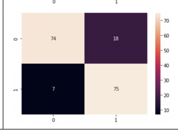
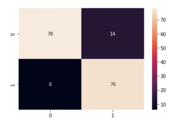
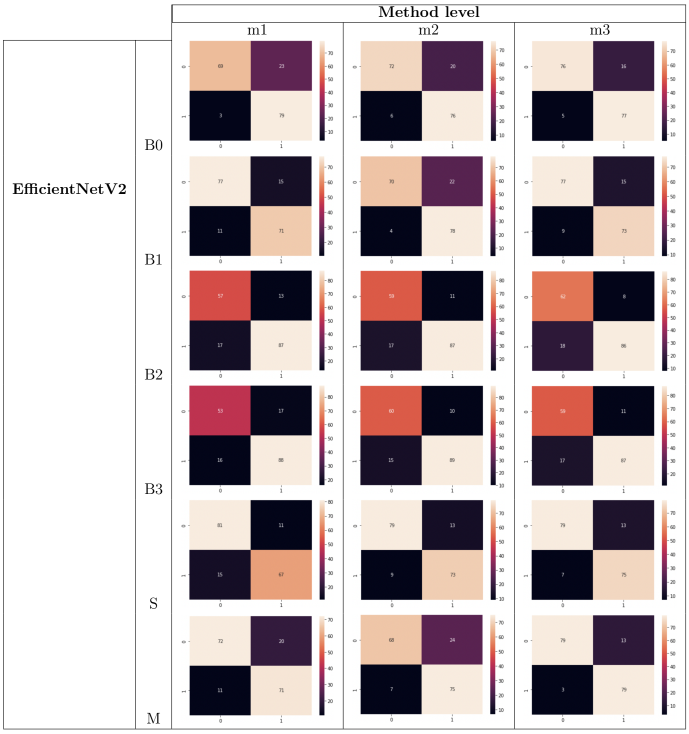
		Method level		
		m1	m2	m3
EfficientNet	B0			
	B1			
	B2			
	B3			
	B4			
	B5			

Table 7.9: Three-stage training system. Confusion Matrix EfficientNetV2 in Fold 0.



The chapter uses the F1-score (higher is better), which is calculated as the harmonic mean between precision and recall, to choose among models. We can observe the increase in performance that the training in three steps confers to the design of the system, very much irrespective of the hyperparameters chosen (learning rate, number of epochs and optimizer).

Table 7.10: Three-stage training system for several baseline models. F1-score, number of trainable parameters and number of non-trainable parameters in Fold 0.

			Method level		
			m1	m2	m3
Model	VGG16	F1-score (%)	84	88	76
		# trainable parameters	2.050	14.678.018	14.716.738
		# non-trainable parameters	14.715.712	39.744	1.024
	InceptionV3	F1-score (%)	78	82	87
		# trainable parameters	8.194	401.410	21.776.546
		# non-trainable parameters	21.806.880	21.413.664	38.528
	ResNet50	F1-score (%)	80	83	89
		# trainable parameters	8.194	5.518.338	23.542.786
		# non-trainable parameters	23.591.808	18.081.664	57.216

Table 7.11: Three-stage training system for several variants of EfficientNet. F1-score, number of trainable parameters and number of non-trainable parameters in Fold 0.

			Method level		
			m1	m2	m3
EfficientNet	B0	F1-score (%)	79	83	88
		# trainable parameters	5.122	1.126.706	4.012.670
		# non-trainable parameters	4.052.131	2.930.547	44.583
	B1	F1-score (%)	81	85	92
		# trainable parameters	5.122	1.355.602	6.518.306
		# non-trainable parameters	6.577.799	5.227.319	64.615
	B2	F1-score (%)	81	85	92
		# trainable parameters	5.634	1.637.594	7.706.628
		# non-trainable parameters	7.771.385	6.139.425	70.391
	B3	F1-score (%)	81	84	87
		# trainable parameters	6.146	1.946.210	10.702.378
		# non-trainable parameters	10.786.607	8.846.543	90.375
	B4	F1-score (%)	80	81	91
		# trainable parameters	7.170	2.643.314	17.555.786
		# non-trainable parameters	17.677.407	15.041.263	128.791
	B5	F1-score (%)	80	86	89
		# trainable parameters	8.194	3.446.914	28.348.978
		# non-trainable parameters	28.517.623	25.078.903	176.839

Table 7.12: Three-stage training system for several variants of EfficientNetV2. F1-score, number of trainable parameters and number of non-trainable parameters in Fold 0.

			Method level		
			m1	m2	m3
EfficientNetV2	B0	F1-score (%)	85	85	88
		# trainable parameters	5.122	594.226	6.865.174
		# non-trainable parameters	6.933.684	6.344.580	73.632
	B1	F1-score (%)	85	85	86
		# trainable parameters	5.122	594.226	6.865.174
		# non-trainable parameters	6.933.684	6.344.580	73.632
	B2	F1-score (%)	82	83	85
		# trainable parameters	5.634	700.406	8.692.720
		# non-trainable parameters	8.772.190	8.077.418	85.104
	B3	F1-score (%)	80	85	83
		# trainable parameters	6.146	860.892	12.827.552
		# non-trainable parameters	12.933.694	12.078.948	112.288
	S	F1-score (%)	85	87	88
		# trainable parameters	5.122	938.050	20.182.610
		# non-trainable parameters	20.333.920	19.400.992	156.432
	M	F1-score (%)	82	82	91
		# trainable parameters	5.122	3.050.626	52.863.478
		# non-trainable parameters	53.152.948	50.107.444	294.592

In addition, transfer learning from Imagenet allows us to rapidly fine-tune the architecture in three stages, achieving high accuracy with limited training time.

The system is implemented using Keras with the following hyperparameters:

- m1 (lr $= 1e-2$, dropout $= 0.2$, epochs $= 50$)
- m2 (lr $= 1e-4$, epochs $= 10$)

- m3 (lr $= 1e-4$, epochs $= 30$)

where 'adam' is the choice of optimizer and the GPU used in the experiments is a Tesla V100 SXM2 (16 GB).

Once the desired model is obtained in Fold 0, we pursue a thorough testing across folds (10-fold crossvalidation) to choose the weights that give better accuracy on a given test subset, see Figure 7.5. In particular, we evaluate the mean and the standard deviation to determine statistical significance of the result.

For the evaluation, we perform 10-fold crossvalidation loading the weights from m3 and retraining on each fold with epochs $= 30$.

7.3.3 Evaluation and Discussion

The chapter builds on VGG16, InceptionV3 and ResNet50 as baseline models of the methodology, and then propose to use variants of EfficientNet to achieve state-of-the-art performance.

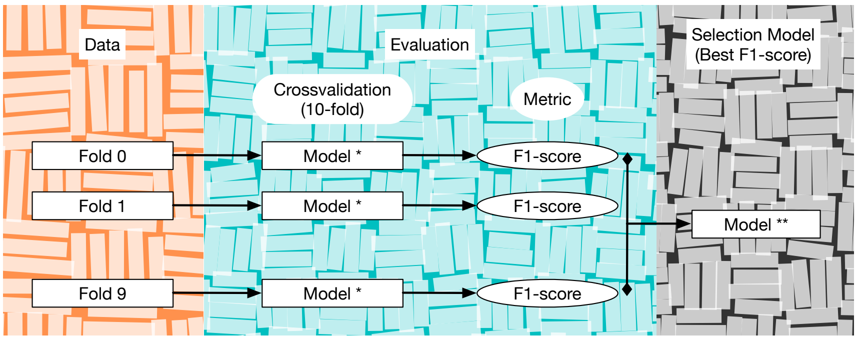


Figure 7.5: Visual description of the evaluation. The weights obtained in the preliminary stage using Fold 0 (Model*: obtained from m3) are then used in 10-fold crossvalidation to retrain the network for each fold, and select the best weights of the model (Model **) according to F1-score.

Extensive evaluation of every model across folds is performed. Tables 7.13, 7.14 and 7.15 show the F1-score across all folds of the dataset evaluating the method under consideration using as initial weights the corresponding weights of m3, that is, the result of the three-stage training in one fold, for each given model. Best results are highlighted, showing the statistical significance of the outputs by computing the mean and standard deviation along the folds.

Plots of accuracy of every model on the sets of training and validation for each fold are shown in Tables 7.16, 7.17 and 7.18 in order to visualize the level of generalization of the architecture.

Table 7.13: Evaluation of the F1-score: baseline models of the method consisting on VGG16, InceptionV3 and ResNet50. Thorough testing across folds with mean and standard deviation for the F1-score of all models under evaluation.

		Fold										**Statistics**	
		0	1	2	3	4	5	6	7	8	9	Mean	stdev
Baseline models	VGG16	83	88	89	83	95	93	60	90	86	65	83,2	1,1
	InceptionV3	91	94	86	92	93	90	91	94	89	92	91,1	2,4
	ResNet50	88	88	82	87	92	85	92	88	91	93	88,9	3,3

Table 7.14: Evaluation of the F1-score (%): methods based on EfficientNet. Thorough testing across folds with mean and standard deviation for the F1-score of all models under evaluation.

		Fold										**Statistics**	
		0	1	2	3	4	5	6	7	8	9	Mean	stdev
EfficientNet	B0	99	99	97	86	97	95	98	98	98	87	96,6	3,7
	B1	92	99	98	94	98	98	97	98	94	93	95,9	2,4
	B2	89	100	96	97	92	98	97	97	98	96	96,1	2,9
	B3	89	99	98	94	98	92	95	98	96	96	95,5	3,0
	B4	97	98	97	98	98	98	96	97	98	91	96,6	2,0
	B5	91	98	97	94	94	96	97	98	94	99	95,8	2,3

Table 7.15: Evaluation of the F1-score (%): methods based on EfficientNetV2. Thorough testing across folds with mean and standard deviation for the F1-score of all models under evaluation.

		Fold										**Statistics**	
		0	1	2	3	4	5	6	7	8	9	Mean	stdev
EfficientNetV2	B0	86	95	96	97	95	93	94	96	93	97	94,3	3,3
	B1	88	95	92	96	96	93	96	96	95	95	94,2	2,5
	B2	94	99	91	95	94	95	95	98	95	87	94,5	3,1
	B3	95	98	96	96	96	96	97	98	95	89	95,7	2,3
	S	93	97	95	97	95	97	97	98	93	93	95,4	1,6
	M	88	97	92	89	92	94	91	98	93	91	92,4	3,0

The three-stage system presented, including variants of both EfficientNet and EfficientNetV2, considerably outperforms the given baselines (VGG16, InceptionV3 and ResNet50), which are similar in scope to the models reported in [6] but using the three-stage training introduced in the chapter. In the case of the baseline models, InceptionV3 has clearly the highest mean F1-score (91.1) compared to VGG16 (83.2) and ResNet50 (88.9). Although InceptionV3 and ResNet50 show comparative performance in terms of number of trained parameters and overall accuracy achieved, the first is more effective with the problem at hand considering that we are dealing with a dataset in the order of the thousands. Should the data to train be increased, it would be expected that ResNet50 achieves slightly better performance due to its better handling of the gradient backpropagating through the layers.

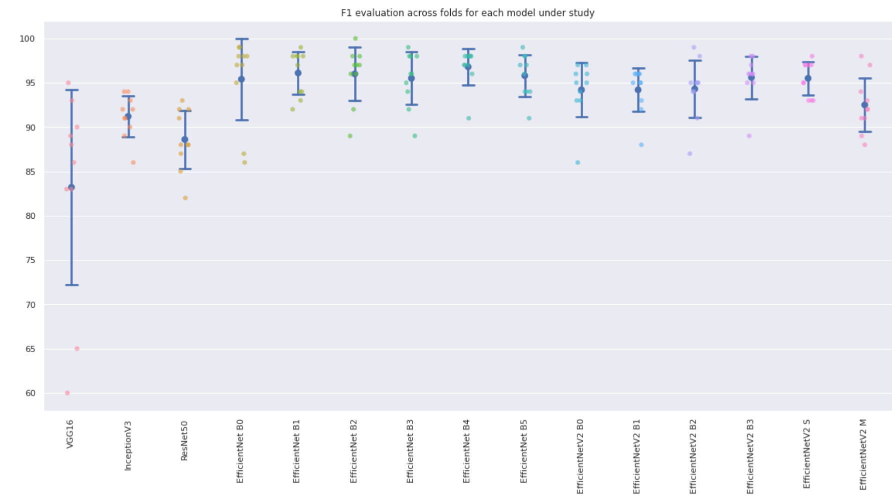


Figure 7.6: F1 evaluation across folds for each model under study using the three-stage training procedure. Error bars with mean and standard deviation for each model are depicted. All architectures based on EfficientNet and EfficientNetV2 outperform the baseline methods (VGG16, InceptionV3 and ResNet50) being EfficientNet B4 and EfficientNetV2 S the best performing techniques.

EfficientNet models perform slightly better in terms of F1-score than EfficientNetV2, although variants of EfficientNetV2 show a better standard deviation across folds; being B2 the model that achieves higher accuracy on a given fold (100 in Fold 1), and B0 and B4 the ones that achieve best mean F1 across folds (96.6), where B4 has the lowest standard deviation (2.0); thus, a better generalization is achieved since the results are more consistent. In the case of EfficientNetV2, V2-B2 achieves the highest accuracy on a given fold (99 in Fold 1), while V2-B3 is the model that gets best mean F1 score across folds (95.7), and model S is the more consistent model according to the standard deviation of the F1-score (1.6). Error bars with mean and standard deviation are showed in Figure 7.6.

Table 7.16: Evaluation on several model baselines (VGG16, InceptionV3 and ResNet50). Accuracy across folds (from 0 to 9).

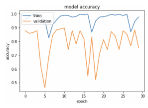
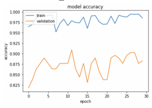
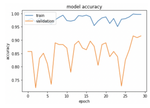
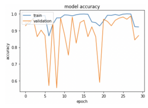
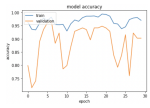
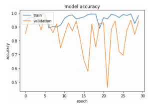
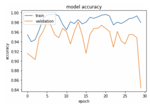
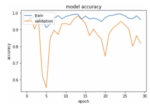
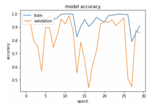
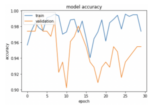
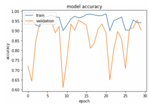
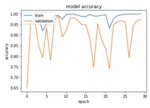
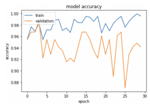
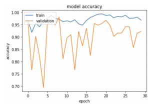
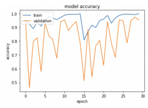
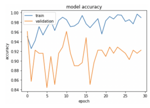
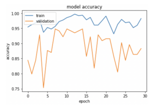
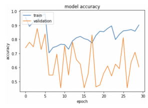
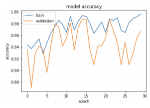
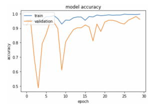
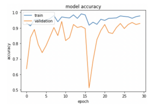
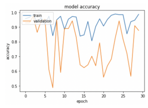
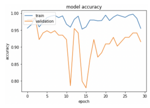
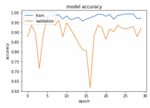
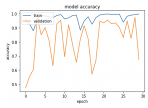
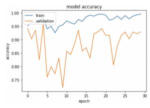
	Baseline models		
Fold	VGG16	InceptionV3	ResNet50
0			
1			
2			
3			
4			
5			
6			
7			
8			
9			

Table 7.17: Evaluation on several variants of EfficientNet (B0-5). Accuracy across folds (from 0 to 9).

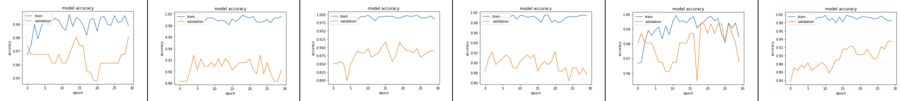
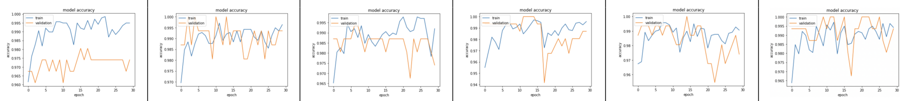
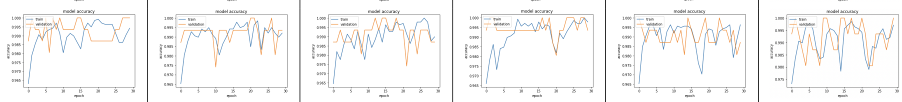
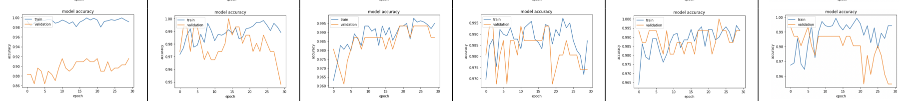
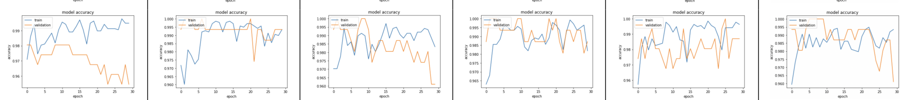
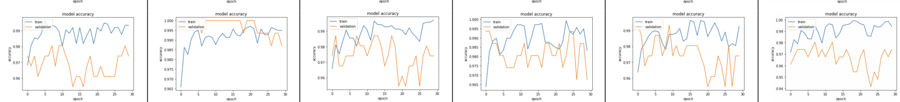
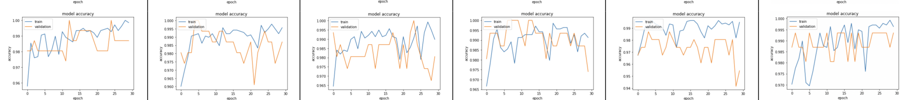
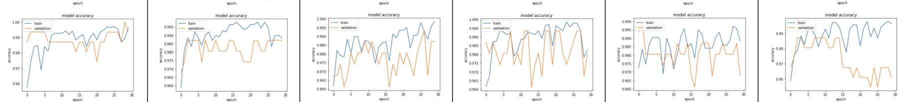
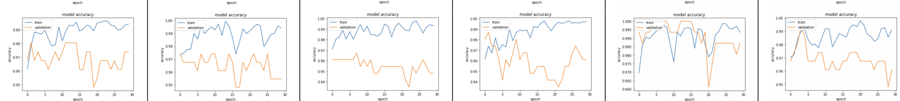
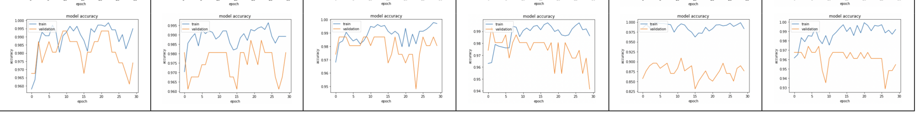
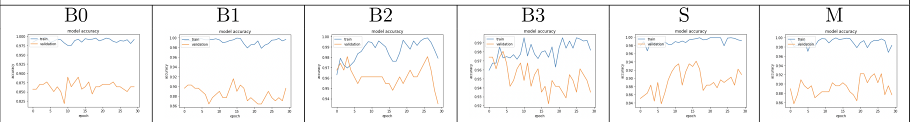
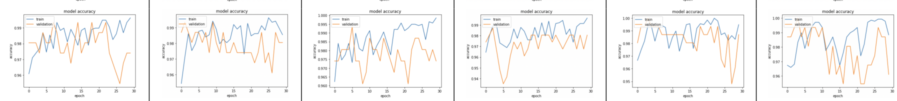
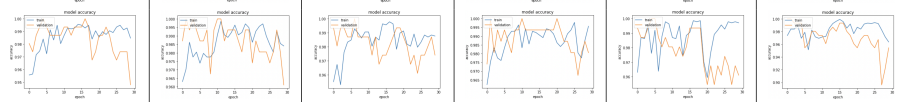
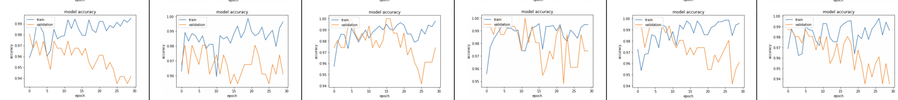
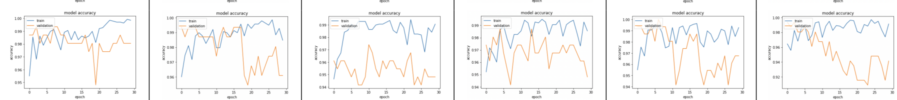
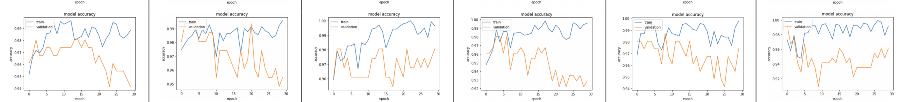
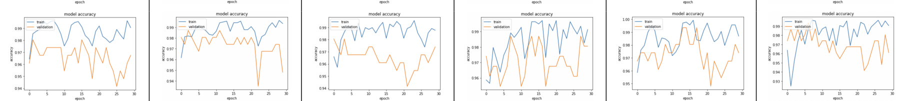
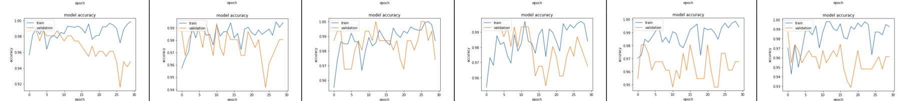
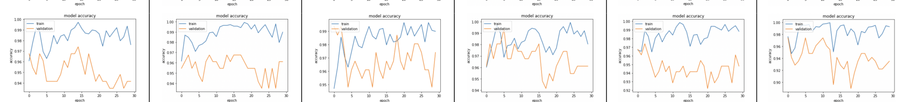
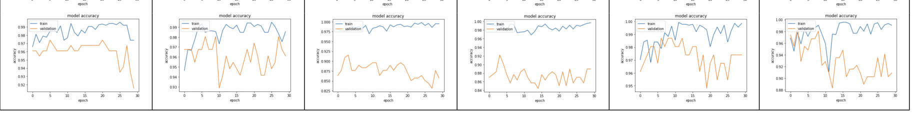
	EfficientNet					
Fold	B0	B1	B2	B3	B4	B5
0						
1						
2						
3						
4						
5						
6						
7						
8						
9						

Table 7.18: Evaluation on several variants of EfficientNetV2 (B0-3, S and M). Accuracy across folds (from 0 to 9).

	EfficientNetV2					
Fold	B0	B1	B2	B3	S	M
0						
1						
2						
3						
4						
5						
6						
7						
8						
9						

Accuracy plots across folds show the considerably good ability to generalize of each model, showing the corresponding curves for the sets of training and validation.

The methodology to tackle the problem in many train stages of the same architecture presents a robust behavior with state-of-the-art performance. Limitations of the technique mainly are due to EfficientNet and EfficientNetV2, where we inherit the necessity to train a large number of parameters, that is clearly less than other state-of-the-art CNN architectures, as the networks are found using NAS optimizing for overall FLOPS, but still very high compared to traditional hand-crafted methodologies where the number of parameters to learn is very low.

7.4 Conclusions and further work

In this work, an intelligent system to automatically detect glaucoma is presented. The methodology is based on a three-stage training procedure based on variants of EfficientNet, a recently proposed family of architectures found using NAS that achieves compelling accuracy on Imagenet, achieving consistent results that outperform the baseline methods. Transfer Learning from Imagenet to the given application under study is employed. The training mechanism applied bestows the system with robustness against hyperparameter choices. We use a dataset consisting of 17.070 fundus images, a considerable size compared to the number of samples used in other recent works, and where the sets used for training, validation and testing are well balanced; such fact confers the obtained models with a low number of false negatives, which is clearly desirable given the gravity and irreversibility of the pathology. Extensive evaluations are reported at each stage of the described procedure under study, as well as, visual interpretation of the results for the sets of training and validation. The F1-score in the test set is used as the target score metric to choose among models, along with a classification report and confusion matrix for each model in the preliminary stage. The proposed system is reliable, highly-accurate, consistent and resource-efficient.

The methodology achieves a mean average percentage F1-score across folds of 96.6 using EfficientNet-B0 (with standard deviation of 3.7) and EfficientNet-B4 (with standard deviation of 2.0), where the best F1 on a given fold is 99 on B0 and 98 on B4. For the case of EfficientNetV2, V2-B3 achieves a mean average F1-score across folds of 95.7 (with standard deviation of 2.3) and V2-S of 95.4 (with standard deviation of 1.6), where the best F1 on a given fold is 98 for both V2-B3 and V2-S. These results significantly outperform the baselines; VGG16 (83.2), InceptionV3 (91.1) and ResNet50 (88.9), and are also clearly better than the state-of-the-art reported results found in the literature, [6].

The three stage-training mechanism using variants of EfficientNet and EfficientNetV2 proposed, although targeted for the particular application of detecting the pathology of glaucoma, achieves a superior classification baseline to use in other clinical conditions, or in the more general case in any vision application where extracting features from raw pixel intensities can play an important role. Indeed, visual sensors are ubiquitous in many applications, such as self-driving cars or Unmanned Aerial Vehicles, where the use of transfer learning, and subsequent freeze, training and finetuning has proven to be very effective; therefore, the system proposed could be further integrated into the detection pipeline of such a system, for instance for lane detection in a self-driving vehicle, or for target recognition in drones.

Bibliography

[1] Barros, D., Moura, J., Freire, C., Taleb, A., Valentim, R., & Morais, P. (2020). Machine learning applied to retinal image processing for glaucoma detection: Review and perspective. *BioMedical Engineering OnLine*, *19*. doi:10.1186/s12938-020-00767-2.

[2] Behrad, F., & Saniee Abadeh, M. (2022). An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications*, *200*, 117006. doi:https://doi.org/10.1016/j.eswa.2022.117006.

[3] Chakravarty, A., & Sivaswamy, J. (2016). Glaucoma classification with a fusion of segmentation and image-based features. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (pp. 689–692). doi:10.1109/ISBI.2016.7493360.

[4] de Curtò, J., de Zarzà, I., Yan, H., & Calafate, C. T. (2022). On the applicability of the hadamard as an input modulator for problems of classification. *Software Impacts*, *13*, 100325. doi:https://doi.org/10.1016/j.simpa.2022.100325.

[5] Diaz-Pinto, A., Colomer, A., Naranjo, V., Morales, S., Xu, Y., & Frangi, A. F. (2019a). Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, *38*, 2211–2218. doi:10.1109/TMI.2019.2903434.

[6] Diaz-Pinto, A., Morales, S., Naranjo, V., Kohler, T., Mossi, J. M., & Navea, A. (2019b). CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *BioMedical Engineering OnLine*, .

[7] Geetha Ramani, R., Balasubramanian, L., & Jacob, S. G. (2012). Automatic prediction of diabetic retinopathy and glaucoma through retinal image analysis and data mining techniques. In *2012 International Conference on Machine Vision and Image Processing (MVIP)* (pp. 149–152). doi:10.1109/MVIP.2012.6428782.

[8] Guo, F., Li, W., Tang, J., Zou, B., & Fan, Z. (2020). Automated glaucoma screening method based on image segmentation and feature extraction. *Medical and Biological Engineering and Computing*, *58*. doi:10.1007/s11517-020-02237-2.

[9] Gupta, I. K., Choubey, A., & Choubey, S. (2022a). Mayfly optimization with deep learning enabled retinal fundus image classification model. *Computers and Electrical Engineering*, *102*, 108176. doi:https://doi.org/10.1016/j.compeleceng.2022.108176.

[10] Gupta, N., Garg, H., & Agarwal, R. (2022b). A robust framework for glaucoma detection using clahe and efficientnet. *Vis. Comput.*, *38*, 2315–2328.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*, .

[12] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *CVPR*, .

[13] Huang, M., Feng, C., Li, W., & Zhao, D. (2021). Vessel enhancement using multi-scale space-intensity domain fusion adaptive filtering. *Biomedical Signal Processing and Control*, *69*, 102799.

[14] Islam, M. T., Mashfu, S. T., Faisal, A., Siam, S. C., Naheen, I. T., & Khan, R. (2022). Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access*, *10*, 2828–2841. doi:10.1109/ACCESS.2021.3139160.

[15] Jaiswal, A. K., Tiwari, P., Kumar, S., Al-Rakhami, M. S., Alrashoud, M., & Ghoneim, A. (2021). Deep learning-based smart iot health system for blindness detection using retina images. *IEEE Access*, *9*, 70606–70615. doi:10.1109/ACCESS.2021.3078241.

[16] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

[17] Maqsood, S., Damaševičius, R., & Maskeliūnas, R. (2021). Hemorrhage detection based on 3d cnn deep learning framework and feature fusion for evaluating retinal abnormality in diabetic patients. *Sensors (Basel)*, .

[18] Mitra, A., Banerjee, P. S., Roy, S., Roy, S., & Setua, S. K. (2018). The region of interest localization for glaucoma analysis from retinal fundus image using deep learning. *Computer Methods and Programs in Biomedicine*, *165*, 25–35. doi:`https://doi.org/10.1016/j.cmpb.2018.08.003`.

[19] Nawaz, M., Nazir, T., Javed, A., Tariq, U., Yong, H.-S., Khan, M. A., & Cha, J. (2022). An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization. *Sensors*, *22*. doi:`10.3390/s22020434`.

[20] Pham, H., Guan, M., Zoph, B., Le, Q., & Dean, J. (2018). Efficient neural architecture search via parameters sharing. *ICML*, .

[21] Rajinikanth, V., Kadry, S., Damaševičius, R., Taniar, D., & Rauf, H. T. (2021). Machine-learning-scheme to detect choroidal-neovascularization in retinal oct image. In *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)* (pp. 1–5). doi:`10.1109/ICBSII51839.2021.9445134`.

[22] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. *MICCAI*, .

[23] Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CVPR*, .

[24] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR*, .

[25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *CVPR*, .

[26] Tan, M., Chen, B., Pang, R., Vasudevan, V., & Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. *CVPR*, .

[27] Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, .

[28] Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. *ICML*, .

[29] Wang, J., Yang, L., Huo, Z., He, W., & Luo, J. (2020). Multi-label classification of fundus images with efficientnet. *IEEE Access*, *8*, 212499–212508. doi:10.1109/ACCESS.2020.3040275.

[30] Wang, X., Xu, M., Li, L., Wang, Z., & Guan, Z. (2019). Pathology-aware deep network visualization and its application in glaucoma image synthesis. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (pp. 423–431). Springer.

[31] Wong, D. W. K., Liu, J., Tan, N.-M., Yin, F., Cheng, X., Cheng, C.-Y., Cheung, G. C. M., & Wong, T. Y. (2012). Automatic detection of the macula in retinal fundus images using seeded mode tracking approach. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, .

[32] Xiong, L., Li, H., & Zheng, Y. (2014). Automatic detection of glaucoma in retinal images. In *2014 9th IEEE Conference on Industrial Electronics and Applications* (pp. 1016–1019). doi:10.1109/ICIEA.2014.6931312.

[33] Yang, J., Huang, M., Fu, J., Lou, C., & Feng, C. (2020a). Frangi based multi-scale level sets for retinal vascular segmentation. *Computer Methods and Programs in Biomedicine*, *197*, 105752.

[34] Yang, J., Lou, C., Fu, J., & Feng, C. (2020b). Vessel segmentation using multiscale vessel enhancement and a region based level set model. *Computerized Medical Imaging and Graphics*, *85*, 101783.

[35] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *4th International Workshop DLMIA* (pp. 3–11). volume 11045. doi:10.1007/978-3-030-00889-5_1.

[36] Zoph, B., & Le, Q. (2017). Neural architecture search with reinforcement learning. *ICLR*, .

[37] Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. *CVPR*, .

Chapter 8

Discussion

In the realm of modern research, there are few areas as dynamic and promising as the integration of LLMs into various fields. From the presented publications, it is apparent that the inclusion of LLMs and the innovations in the field of AI [35, 36] are significantly shaping and pushing the boundaries in areas such as truck platooning [37, 38], multi-agent systems [40, 39, 41], UAV scene understanding [12, 27], forest fire estimation [44], and medical diagnostics [45].

B5G Truck Platooning Systems: The use of LLMs in B5G truck platooning systems promises to make transportation not only more efficient but also safer. The authors illustrate the potential of LLMs in a telecommunication environment, presenting an ambitious avenue for AI-driven control in vehicular systems. The discussion on the future of LLMs hints at a world where these models are deeply embedded in safety-critical domains beyond just transportation.

Multi-Agent Systems and Evolutionary Coevolution (EC) Theory [42, 43]: The blend of EC Theory and MASs augmented by LLMs reveals a more profound understanding of cooperation, competition, and defection in complex environments. The integration of LLMs into this matrix serves as an innovative form of human-computer interface, presenting a clear trajectory towards more efficient and effective strategies within Human-Computer Interaction.

UAV Semantic Scene Understanding: Drones, or UAVs, have rapidly become pivotal in various sectors. The research introduces a fascinating perspective of integrating LLMs with UAVs for scene understanding. Such methodologies can revolutionize surveillance, reconnaissance, and potentially even rescue missions, emphasizing the versatility of LLMs in practical applications.

Forest Fire Area Estimation: The devastation caused by forest fires worldwide underscores the need for advanced systems to predict and manage them. The utilization of advanced ML and DL techniques, like the TabNet combined with transformers, offers promising avenues for rapid and accurate estimations, potentially saving billions in damages and countless lives.

Glaucoma Detection: In the medical domain, early detection of conditions like glaucoma can drastically change the prognosis for patients. The three-stage training methodology using EfficientNet variants sets a benchmark not just for medical diagnostics but broadly for any vision application. Such advancements make the goal of reliable, rapid, and affordable diagnostic systems a tangible reality.

Across all these studies, a recurring theme is the marriage of traditional techniques with state-of-the-art AI methodologies. For instance, traditional PID controls, well-understood theories from game theory and MASs, established UAV operations, known statistical techniques for fire estimation, and existing medical imaging methods are all enhanced exponentially by integrating them with advanced AI strategies and LLMs. Such a combination does not merely add to the existing capabilities but often multiplies them, presenting significant breakthroughs.

In summation, the presented publications reflect a paradigm shift in the application of LLMs across diverse fields [46]. Each study not only showcases the profound impact of LLMs and DL but also sets a roadmap for future research, hinting at an era where the boundaries between human intelligence and AI become increasingly blurred.

From safer roads with B5G truck platooning systems to a deeper understanding of cooperative behaviors in MASs, from advanced UAVs capable of nuanced scene understanding to accurate estimations of forest fire devastation, and from rapid medical diagnostics to potential applications in other vision tasks, the research underscores the sheer breadth and depth of possibilities with LLMs and sophisticated learning techniques.

The journey of integrating LLMs into these domains has only just begun. As technology continues to evolve, it is evident that the confluence of traditional methods with advanced AI and LLMs will lead to innovative solutions, making our world safer, more efficient, and more interconnected.

Table 8.1 provides a comparative analysis of these chapters, outlining the techniques employed, key findings, potential applications, and directions for future research. This summary serves as a concise guide, offering readers a panoramic view of the topics covered in the doctoral dissertation.

As a summary, throughout the chapters presented, there is a recurring theme of harnessing the capabilities of LLMs and integrating them with specialized domains, revealing both their immense potential and the associated challenges. Chapter 3 underlines the synergy between traditional control systems and the dawn of AI, illuminating the transformative capabilities of LLMs in steering complex autonomous systems. Meanwhile, Chapter 4 offers a profound exploration into the nuanced dance of cooperation and defection, emphasizing the need to reconcile academic theory with real-world intricacies. Chapter 5 pioneers the realm of semantic interpretation, championing the adaptability of human-computer interactions, even while acknowledging LLM limitations. Chapter 6 gravitates towards a pressing environmental concern, showcasing the promise and requisites of ML in estimating forest fire impacts. Finally, Chapter 7 stands as a testament to the strides in medical diagnostics through DL, with the versatility of EfficientNet promising far-reaching applications. Collectively, these chapters not only contribute to their respective fields but also illuminate the road ahead, spotlighting areas of potential breakthroughs and the inherent challenges awaiting advancements.

Table 8.1: Comparative Analysis of Chapters 3 to 7

Chapter	Techniques & Methodologies Used	Main Contributions/Findings	Potential Applications	Future Research Directions
Chapter 3	Control systems, LLMs (e.g., GPT-3.5-turbo)	Evolution of control systems with LLMs	Engineering domains	Real-world data validation for AI-driven control
Chapter 4	EC Theory, MASs, LLMs	Implications of EC in business and societal contexts	Business and societal contexts	Addressing complexities inherent in systems
Chapter 5	UAVs, LLM pipeline	Socratic video understandings	Human-computer interaction	Probabilistic reasoning and visual data interpretation
Chapter 6	ML, DL, TabNet	Estimating forest fire impact	Predictive models for forest fires	Model performance on diverse datasets
Chapter 7	variants of EfficientNet	High accuracy glaucoma detection	Medical domain, self-driving cars, drones	Transfer learning and integration into diverse applications

8.1 Contributions

In a comprehensive effort to understand and harness advanced ML models for diverse applications, this doctoral thesis presents empirical insights derived from rigorously designed experiments. Within the landscape of B5G truck platooning, our exploration into adaptive PID control, reinforced by LLMs, especially the GPT-3.5-turbo, showed promising enhancements in reliability and security parameters of vehicle coordination in technologically evolving communication environments. As we transitioned into examining Multi-Agent Systems, our EC Theory illuminated previously uncharted dynamics among heterogeneous agents. This was further corroborated by our simulation results which visualized the emergence of cooperative behavior patterns, thus underscoring LLMs' potential in facilitating cooperation and resilience in such systems. Socratic video understanding, as applied in UAVs, achieved a significant linguistic complexity with generated texts reaching a GUNNING Fog median grade level between 7-12. This prowess, combined with the sophisticated collaboration between BLIP-2 and GPT-3.5-turbo, brought forth actionable insights from aerial video footage. In our forest fire area estimation study, the TabNet transformer-based architecture outperformed traditional methods with a marked reduction in error, showcasing a MSE of 2319 in training and 7781 in testing. Lastly, our efforts in detecting glaucoma using EfficientNet highlighted the robustness of a three-stage-training methodology. By iteratively fine-tuning the model, we achieved unmatched precision in glaucoma detection, as evidenced by our accuracy and F1-score metrics, while simultaneously maintaining a lower parameter footprint compared to conventional models like VGG16 and Resnet50. Collectively, these findings not only substantiate the value proposition of our research endeavors but also underscore the overarching theme: the profound influence of LLMs and advanced DL architectures in shaping the future of intelligent systems across diverse domains.

Chapter 9

Conclusions

9.1 Concluding Remarks

In a world that constantly experiences an upsurge of technological advancements, research efforts such as those presented in the highlighted papers continue to widen the horizons of both academic and practical applications. By integrating advanced computational methodologies and AI, these studies are poised to revolutionize their respective domains.

Collectively, these studies encapsulate the power and potential of intertwining advanced algorithms, AI, and practical applications. Each paper, while addressing a distinct domain, underscores the centrality of AI in resolving real-world challenges.

As technology continues its relentless march forward, these foundational research efforts not only provide insight into current challenges, but also outline the trajectory of future explorations. Whether it is enhancing transportation through intelligent truck platooning, fostering human-AI collaboration, optimizing UAV applications, managing natural calamities, or revolutionizing healthcare diagnostics, the road ahead promises innovation, integration, and improvement.

In the grand tapestry of scientific endeavor, these researches are emblematic threads, each contributing to the larger understanding of our world and the technologies that shape it. We eagerly anticipate future endeavors building on these findings, ultimately advancing humanity's collective knowledge and capabilities.

Throughout this thesis, we have delicately navigated the intricate landscapes of LLMs, advanced DL architectures, and their applications, employing both theoretical constructs and empirical validations. The mathematical robustness underlying our methodologies, from the adaptive mechanisms in PID controllers to the nuanced complexities in the EC Theory, provides a strong foundation for further analytical inquiries. One potential avenue of exploration is the formalization of LLM-enhanced dynamics in multi-agent systems using differential game theory, $\mathcal{G}(\mathcal{N}, \mathcal{S}, \mathcal{U}, f)$, where $\mathcal{N}$ represents the set of agents, $\mathcal{S}$ their state spaces, $\mathcal{U}$ their action profiles, and f the state transitions influenced by LLM insights. This could provide deeper insights into cooperative and non-cooperative equilibria, $\varepsilon(\mathcal{N}, \mathcal{U})$, especially in systems where agent strategies are continually evolving under the influence of LLM-based recommendations. Additionally, the amalgamation of transformer architectures like TabNet with topological data analysis could pave the way for understanding high-dimensional, non-linear data structures intrinsic to natural phenomena, such as forest fires. With EfficientNet's demonstrated prowess, the adaptation of its architectural nuances to other domains, perhaps coupled with variational methods, $\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z))]$, could unlock unprecedented diagnostic efficiencies. As we stand on the precipice of a new era in AI and DL, the foundational work of this thesis serves not as a culmination, but as a beacon, heralding the boundless explorations and innovations yet to come.

9.2 Publications Included in the Thesis

The following publications, listed below, have been included in this thesis and represent its core contributions:

1. [47] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. (2023). "LLM Adaptive PID Control for B5G Truck Platooning Systems." *Sensors*,

vol(23), 5899. DOI: 10.3390/s23135899 IF 2022: 3.9; SCImago-SJR: 1r cuartil.

2. [48] de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P., & Calafate, C. T. (2023). "Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs." *Electronics*, vol(12), 2722. DOI: 10.3390/electronics12122722 IF 2022: 2.9; JCR: 2o cuartil.

3. [49] de Zarzà, I., de Curtò, J., Calafate, C. T. (2023). "Socratic Video Understanding on Unmanned Aerial Vehicles." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pending assignment. CORE B.

4. [50] de Zarzà, I., de Curtò, J., Calafate, C. T. (2023). "Area Estimation of Forest Fires using TabNet with Transformers." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pending assignment. CORE B.

5. [51] de Zarzà, I., de Curtò, J., Calafate, C. T. (2022). "Detection of glaucoma using three-stage training with EfficientNet." *Intelligent Systems with Applications*, vol(16), 200140.
DOI: 10.1016/j.iswa.2022.200140 SCImago-SJR: 1r cuartil.

9.3 Related Publications

In addition to the main publications, the following works have been conducted in relation to the broader research context:

1. [52] de Zarzà, I., de Curtò, J., Cano, J. C., & Calafate, C. T. (2023) "Drone-Based Decentralized Truck Platooning with UWB Sensing and Control" Mathematics, vol(11), 4627.
DOI: 10.3390/math11224627 IF 2022: 2.4; JCR: 1r cuartil.

2. [53] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. (2023). "LLM Multimodal Traffic Accident Forecasting" Sensors, vol(23), 9225.
DOI: 10.3390/s23229225. IF 2022: 3.9; SCImago-SJR: 1r cuartil.

3. [54] de Curtò, J., de Zarzà, I., & Calafate, C. T. (2023). "Semantic Scene Understanding with Large Language Models on Unmanned Aerial

Vehicles." *Drones*, vol(7), 114.
DOI: 10.3390/drones7020114 IF 2022: 4.8; SCImago-SJR: 1r cuartil.

4. [55] de Curtò, J., de Zarzà, I., Roig, G., Manzoni, P., & Calafate, C. T. (2023). "LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments." *Electronics*, vol(12), 2814. DOI: 10.3390/electronics12132814 IF 2022: 2.9; JCR: 2o cuartil.

5. [56] de Zarzà, I., de Curtò, J., Hernández-Orallo, E., & Calafate, C. T. (2023). "Cascading and Ensemble Techniques in Deep Learning." *Electronics*, vol(12), 3354. DOI: 10.3390/electronics12153354 IF 2022: 2.9; JCR: 2o cuartil.

6. [57] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "Optimizing Neural Networks for Imbalanced Data." *Electronics*, vol(12), 2674. DOI: 10.3390/electronics12122674 IF 2022: 2.9; JCR: 2o cuartil.

7. [58] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. (2023). "Signature and Log-Signature for the Study of Empirical Distributions Generated with GANs." *Electronics*, vol(12), 2192. DOI: 10.3390/electronics12102192 IF 2022: 2.9; JCR: 2o cuartil.

8. [59] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. (2023). "Summarization of Videos with the Signature Transform." *Electronics*, vol(12), 1735. DOI: 10.3390/electronics12071735 IF 2022: 2.9; JCR: 2o cuartil.

9. [60] de Zarzà, I., de Curtò, J., & Calafate, C. T. "UMAP for Geospatial Data Visualization." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pending assignment. CORE B.

10. [61] de Zarzà, I., de Curtò, J., & Calafate, C. T. "Decentralized Platooning Optimization for Trucks: A MILP and ADMM-based Convex Approach to Minimize Latency and Energy Consumption." 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023), Hong Kong. July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00031. CORE A.

11. [62] de Zarzà, I., de Curtò, J., & Calafate, C. T. "Decentralized Planning of Platoons in Road Transport using Reinforcement Learning." 6th International Workshop on Vehicular Networking and Intelligent Trans-

portation Systems (VENITS 2023), Hong Kong. July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00030. CORE A.

12. [63] de Curtò, J., de Zarzà, I., & Calafate, C. T. (2023). "UWB and MB-OFDM for Lunar Rover Navigation and Communication." *Mathematics*, vol(11), 3835. DOI: 10.3390/math11183835 IF 2022: 2.4; JCR: 1r cuartil.

13. [64] de Curtò, J., de Zarzà, I., Yan, H., & Calafate, C. T. (2022). "On the applicability of the Hadamard as an input modulator for problems of classification." *Software Impacts*, vol(13), 100325.
DOI: 10.1016/j.simpa.2022.100325 IF 2022: 2.1; JCR: 3r cuartil.

9.4 Open Science and Reproducibility

In line with the principles of open science and to facilitate the verification and reproducibility of our results, representative software examples developed as part of this research has been made publicly available.

9.4.1 Representative Open-Source Code

- **LLM Multimodal Traffic Accident Forecasting [53]:**
  - *DOI:* 10.6084/m9.figshare.24570478
  - *Description:* This repository contains the VLM code, along with statistical and DL techniques used to forecast traffic accidents, demonstrating the application of multimodal AI in traffic accident forecasting in the scenario of autonomous driving. The representative code examples also highlight the application of probabilistic hashing and deep feature extraction.
- **Drone-Based Decentralized Truck Platooning [52]:**
  - *DOI:* 10.6084/m9.figshare.24549
  - *Description:* Available in this repository are the simulation tools used to study truck platooning systems assisted by drones, showcasing the integration of autonomous vehicles and UAVs.

- **Detection of Glaucoma Using EfficientNet [51]:**
  - *DOI:* 10.24433/CO.8342269.v1
  - *Description:* This code repository provides the implementation of the three-stage training mechanism using the EfficientNet model for the detection of glaucoma, emphasizing advances in medical diagnostics.

9.5 Future Work

The research carried out in this thesis has opened several avenues for further exploration, pushing the boundaries of current knowledge and technological capabilities. As we chart our journey forward, the following areas have emerged as promising domains for future endeavors:

1. **Deepening the Coevolutionary Theory:** The introduction of LLMs in the coevolutionary domain offers a new perspective on emergent cooperation. Future work could explore the coadaptation of LLMs in various dynamic environments, leveraging the symbiotic relationship between learning and adaptation.
2. **Granular Control Mechanisms:** The success of the adaptive PID control for truck platooning opens the door to the investigation of more granular control strategies. Incorporating AI-driven predictive modeling could further enhance the responsiveness and efficiency of these systems.
3. **Expanding UAV Capabilities:** The use of LLMs for semantic scene understanding on UAVs signifies a pivotal advancement in drone technology. There is potential to integrate multi-modal sensor fusion, combining visual, thermal, and sonar data, to enhance UAV perception in varying environmental conditions.
4. **Forest Fire Prediction:** The current work on area estimation of forest fires can be augmented with real-time prediction models, possibly using time-series analysis combined with meteorological data. This would be invaluable for preemptive measures and resource allocation during wildfire events.
5. **Medical Imaging and Diagnostics:** The promising results from glaucoma detection point towards a broader horizon in medical diagnostics. Combining LLMs with generative networks could potentially simulate

medical conditions, aiding in training models even with limited real-world data.

In addition to these specific areas, there remains a perennial quest for enhancing computational efficiency, ensuring ethical considerations in AI applications, and bridging the gap between theory and real-world applicability. As we stand at this juncture, the future beckons with a myriad of challenges and opportunities, each promising to push the boundaries of what we perceive as possible today.

9.6 Synthesis of Contributions

The body of work presented in this thesis, substantiated through publications in peer-reviewed journals and conferences, collectively contributes to the field of AI-enhanced autonomous systems. The following section synthesizes these contributions to articulate the "big picture" of this research's impact.

The studies within this thesis represent significant leaps in the realms of control systems, strategic interactions, scene understanding, environmental management, and healthcare diagnostics:

- **Control Systems Enhancement:** The integration of LLMs within vehicular platooning has demonstrated that AI can significantly enhance communication and decision-making efficacy, setting a precedent for future transportation systems.
- **Strategic Interactions in MASs:** By infusing EC Theory with LLM-based insights, we have unveiled new cooperative behaviors, establishing a framework for AI-mediated social and economic interactions.
- **UAV Scene Understanding:** The development of a Socratic understanding model using LLMs and VLMs for UAVs has pushed the envelope in video analysis, enabling nuanced, context-aware interpretations of visual data streams.
- **Environmental Management:** The application of TabNet for forest fire area estimation has proven the potency of DL in environmental applications, providing a model for disaster management and response strategies.
- **Healthcare Diagnostics:** Our work with EfficientNet for glaucoma detection exemplifies the transformative potential of DL in medical imag-

ing, potentially leading to breakthroughs in early diagnosis and treatment strategies.

The convergence of LLMs, DL, and EC within this thesis underscores the synergistic potential of hybrid AI approaches. The integration of these methodologies addresses complex challenges by leveraging their combined strengths, thus offering a multidimensional perspective on problem-solving in autonomous systems.

As we extrapolate from these individual studies, the main narrative that emerges is one of interconnectedness—between diverse AI technologies and between computational innovation and its applications. The contributions of this thesis are therefore not confined to their immediate fields but extend to influencing the trajectory of future research and development:

- **Interdisciplinary Impact:** The methodologies and findings presented have implications that transcend their respective fields, suggesting potential crossover applications and interdisciplinary research opportunities.
- **Future Research Pathways:** The foundational work laid down by this thesis illuminates the pathway for future exploration, including the integration of AI in more complex systems, the refinement of models for greater efficiency and accuracy, and the ethical deployment of AI technologies.

In conclusion, the collective insights gleaned from the research presented in this thesis contribute to a broader understanding of how advanced AI methodologies can be harnessed to address and solve real-world challenges. As we reflect on these contributions, it is evident that they embody a stepping stone towards a future where AI is seamlessly integrated into the fabric of our daily lives, enhancing our capabilities, and expanding the horizons of human potential.

Acronyms

AES	Advanced Encryption Standard
AI	Artificial Intelligence
DC	Drought Code
DL	Deep Learning
DMC	Duff Moisture Code
DNN	Deep Neural Networks
EC	Extended Coevolutionary
FFMC	Fine Fuel Moisture Code
GPT	Generative Pretraining Transformer
HCI	Human–Computer Interaction
ISI	Initial Spread Index
LLM	Large Language Models
MAS	Multi-Agent Systems
ML	Machine Learning
MSE	Mean Squared Error
NAS	Neural Architecture Search
NLP	Natural Language Processing
PID	Proportional-Integral-Derivative
RH	Relative Humidity
SMs	Socratic Models
UAV	Unmanned Aerial Vehicle
V2V	Vehicle-to-Vehicle
VLMs	Visual Language Models

Bibliography

[1] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[2] Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

[3] Muglich, D.; Schroeder de Witt, C.; van der Pol, E.; Whiteson, S.; Foerster, J. Equivariant networks for zero-shot coordination. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6410–6423.

[4] Yang, M.S.; Schuurmans, D.; Abbeel, P.; Nachum, O. Chain of thought imitation with procedure cloning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36366–36381.

[5] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *CVPR*, .

[6] Zoph, B., Le, Q., 2017. Neural architecture search with reinforcement learning. ICLR .

[7] Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. *CVPR*, .

[8] Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; Florence, P. PaLM-E: An Embodied Multimodal Language Model. In *arXiv preprint arXiv:2303.03378* **2023**.

[9] Yang, J.; Peng, W.; Sun, C. A learning control method of automated vehicle platoon at straight path with ddpg-based pid. *Electronics* **2021**, *10*, 2580. [CrossRef]

[10] Shoham, Y.; Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*; Cambridge University Press: Cambridge, UK, 2008.

[11] Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624.

[12] Bonatti, R., Bucker, A., Scherer, S., Mukadam, M., Hodgins, J., 2021. Batteries, camera, action! learning a semantic control space for expressive robot cinematography. ICRA .

[13] Xiong, L., Li, H., & Zheng, Y. (2014). Automatic detection of glaucoma in retinal images. In *2014 9th IEEE Conference on Industrial Electronics and Applications* (pp. 1016–1019). doi:10.1109/ICIEA.2014.6931312.

[14] Chakravarty, A., & Sivaswamy, J. (2016). Glaucoma classification with a fusion of segmentation and image-based features.

[15] Maqsood, S., Damaševičius, R., & Maskeliūnas, R. (2021). Hemorrhage detection based on 3d cnn deep learning framework and feature fusion for evaluating retinal abnormality in diabetic patients. *Sensors (Basel)* .

[16] Boysen, N.; Briskorn, D.; Schwerdfeger, S. The identical-path truck platooning problem. *Transp. Res. Part B Methodol.* **2018**, *109*, 26–39. [CrossRef]

[17] Åström, K.J.; Hägglund, T. *PID Controllers: Theory, Design, and Tuning*; Instrument Society of America: Research Triangle Park, NC, USA 1995.

[18] Åström, K.J.; Wittenmark, B. *Adaptive Control*; Dover Publications: Mineola, NY, USA, 2013.

[19] Franklin, G.F.; Powell, J.D.; Emami-Naeini, A. *Feedback Control of Dynamic Systems*; Prentice Hall: Hoboken, NJ, USA, 2014.

[20] Busoniu, L.; Babuska, R.; Schutter, B.D. A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2008**, *38*, 156–172. [CrossRef]

[21] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019 ; Volume 1, pp. 4171–4186.

[22] Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., Lee, J., Vanhoucke, V., Florence, P., 2022. Socratic models: Composing zero-shot multimodal reasoning with language.

[23] Li, J., Li, D., Xiong, C., Hoi, S., 2022b. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: ICML.

[24] Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv.

[25] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf` (accessed on 26 April 2023).

[26] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. 2019. Available online: `https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf` (accessed on 26 April 2023).

[27] Vemprala, S., Bonatti, R., Bucker, A., Kapoor, A., 2023. ChatGPT for Robotics: Design Principles and Model Abilities. Technical Report MSR-TR-2023-8. Microsoft. URL: `https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/`.

[28] Arik, S.O., Pfister, T., 2021. TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence 35, 6679–6687. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/16826`, doi:`10.1609/aaai.v35i8.16826`.

[29] Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, .

[30] Tan, M., Chen, B., Pang, R., Vasudevan, V., Le, Q.V., 2019. Mnasnet: Platform-aware neural architecture search for mobile. CVPR .

[31] Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. *ICML*, .

[32] Wang, X., Xu, M., Li, L., Wang, Z., & Guan, Z. (2019). Pathology-aware deep network visualization and its application in glaucoma image synthesis. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (pp. 423–431). Springer.

[33] Zeng, A., Florence, P., Tompson, J., Welker, S., Chien, J., Attarian, M., Armstrong, T., Krasin, I., Duong, D., Wahid, A., Sindhwani, V., Lee, J., 2020. Transporter networks: Rearranging the visual world for robotic manipulation. CoRL .

[34] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L., 2022. OPT: Open pre-trained transformer language models. arXiv:2205.01068 .

[35] Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Visual Language Maps for Robot Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023.

[36] Shah, D.; Osiński, B.; Ichter, B.H.; Levine, S. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In Proceedings of The 6th Conference on Robot Learning, Auckland, New Zealand, 14–18 December 2022; Volume 205, pp. 492–504.

[37] Larsson, E.; Sennton, G.; Larson, J. The vehicle platooning problem: Computational complexity and heuristics. *Transp. Res. Part C Emerg. Technol.* **2015**, *60*, 258–277. [CrossRef]

[38] Boubakri, A.; MataliGmmar, S. Platoons formation management strategies based on reinforcement learning. In Proceedings of the Advances in Systems Engineering: Proceedings of the 28th International Conference on Systems Engineering (ICSEng 2021), Wrocław, Poland, 14–16 December 2021; pp. 57–66.

[39] Nash, J. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49. [CrossRef]

[40] von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 1944.

[41] Stone, P.; Kraus, S. To teach or not to teach? Decision making under uncertainty in ad hoc teams. In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, ON, Canada, 10–14 May 2010; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 2010; Volume 1, pp. 157–164.

[42] Gong, D.; Xu, B.; Zhang, Y.; Guo, Y.; Yang, S. A similarity-based cooperative co-evolutionary algorithm for dynamic interval multiobjective optimization problems. *IEEE Trans. Evol. Comput.* **2019**, *24*, 142–156. [CrossRef]

[43] Meselhi, M.A.; Elsayed, S.M.; Sarker, R.A.; Essam, D.L. Contribution based co-evolutionary algorithm for large-scale optimization problems. *IEEE Access* **2020**, *8*, 203369–203381. [CrossRef]

[44] Kurbanov, E., Vorobev, O., Lezhnin, S., Sha, J., Wang, J., Li, X., Cole, J., Dergunov, D., Wang, Y., 2022. Remote sensing of forest burnt area, burn severity, and post-fire recovery: A review. Remote Sensing 14. URL: `https://www.mdpi.com/2072-4292/14/19/4714`, doi:`10.3390/rs14194714`.

[45] Behrad, F., & Saniee Abadeh, M. (2022). An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications*, *200*, 117006. doi:`https://doi.org/10.1016/j.eswa.2022.117006`.

[46] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez,

E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B., 2022. RT-1: Robotics transformer for real-world control at scale, in: arXiv preprint arXiv:2212.06817.

[47] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. LLM Adaptive PID Control for B5G Truck Platooning Systems. *Sensors*, 23:5899, 2023. DOI: 10.3390/s23135899.

[48] de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P., & Calafate, C. T. Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs. *Electronics*, 12:2722, 2023. DOI: 10.3390/electronics12122722.

[49] de Zarzà, I., de Curtò, J., & Calafate, C. T. Socratic Video Understanding on Unmanned Aerial Vehicles. In *Proceedings of the 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023)*, Athens, Greece, 6–8 September, 2023.

[50] de Zarzà, I., de Curtò, J., & Calafate, C. T. Area Estimation of Forest Fires using TabNet with Transformers. In *Proceedings of the 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023)*, Athens, Greece, 6–8 September, 2023.

[51] de Zarzà, I., de Curtò, J., & Calafate, C. T. Detection of glaucoma using three-stage training with EfficientNet. *Intelligent Systems with Applications*, 16:200140, 2022. DOI: 10.1016/j.iswa.2022.200140.

[52] de Zarzà, I.; de Curtò, J.; Cano, J.C.; Calafate, C.T. Drone-Based Decentralized Truck Platooning with UWB Sensing and Control. *Mathematics*, 11:4627, 2023. DOI: 10.3390/math11224627.

[53] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. LLM Multimodal Traffic Accident Forecasting. *Sensors*, 23:9225, 2023. DOI: 10.3390/s23229225.

[54] de Curtò, J., de Zarzà, I., & Calafate, C. T. Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles. *Drones*, 7:114, 2023. DOI: 10.3390/drones7020114.

[55] de Curtò, J., de Zarzà, I., Roig, G., Manzoni, P., & Calafate, C. T. LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments. *Electronics*, 12:2814, 2023. DOI: 10.3390/electronics12132814.

[56] de Zarzà, I., de Curtò, J., Hernández-Orallo, E., & Calafate, C. T. Cascading and Ensemble Techniques in Deep Learning. *Electronics*, 12:3354, 2023. DOI: 10.3390/electronics12153354.

[57] de Zarzà, I., de Curtò, J., & Calafate, C. T. Optimizing Neural Networks for Imbalanced Data. *Electronics*, 12:2674, 2023. DOI: 10.3390/electronics12122674.

[58] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. Signature and Log-Signature for the Study of Empirical Distributions Generated with GANs. *Electronics*, 12:2192, 2023. DOI: 10.3390/electronics12102192.

[59] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. Summarization of Videos with the Signature Transform. *Electronics*, 12:1735, 2023. DOI: 10.3390/electronics12071735.

[60] de Zarzà, I., de Curtò, J., & Calafate, C. T. UMAP for Geospatial Data Visualization. In *Proceedings of the 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023)*, Athens, Greece, 6–8 September, 2023.

[61] de Zarzà, I., de Curtò, J., & Calafate, C. T. Decentralized Platooning Optimization for Trucks: A MILP and ADMM-based Convex Approach to Minimize Latency and Energy Consumption. In *Proceedings of the 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023)*, Hong Kong, July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00031.

[62] de Zarzà, I., de Curtò, J., & Calafate, C. T. Decentralized Planning of Platoons in Road Transport using Reinforcement Learning. In *Proceedings of the 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023)*, Hong Kong, July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00030.

[63] de Curtò, J., de Zarzà, & Calafate, C. T. UWB and MB-OFDM for Lunar Rover Navigation and Communication. *Mathematics*, 11:3835, 2023. DOI: 10.3390/math11183835.

[64] de Curtò, J., de Zarzà, I., Yan, H., & Calafate, C. T. On the applicability of the Hadamard as an input modulator for problems of classification. *Software Impacts*, 13:100325, 2022. DOI: 10.1016/j.simpa.2022.100325.