ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

FUEL CELL AIR SUPPLY SYSTEM CONTROL USING DEEP Q-NETWORK

M.Sc. THESIS

Emir OĞUZ

Department of Control and Automation Engineering

Control and Automation Engineering Programme

JUNE 2026

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

FUEL CELL AIR SUPPLY SYSTEM CONTROL USING DEEP Q-NETWORK

M.Sc. THESIS

Emir OĞUZ
(504231112)

Department of Control and Automation Engineering

Control and Automation Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Ahmet ONAT

JUNE 2026

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DERİN Q-AĞI KULLANILARAK YAKIT HÜCRESİ HAVA BESLEME SİSTEMİ KONTROLÜ

YÜKSEK LİSANS TEZİ

Emir OĞUZ
(504231112)

Kontrol ve Otomasyon Mühendisliği Anabilim Dalı

Kontrol ve Otomasyon Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Ahmet ONAT

HAZİRAN 2026

Emir OĞUZ, a M.Sc. student of İTU Graduate School student ID 504231112, successfully defended the thesis entitled “FUEL CELL AIR SUPPLY SYSTEM CONTROL USING DEEP Q-NETWORK”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assoc. Prof. Dr. Ahmet ONAT** İstanbul Technical University	.....
Jury Members :	**Prof. Dr. Mustafa DOĞAN** İstanbul Technical University	.....
	Prof. Dr. Şeref Naci ENGİN Yıldız Technical University	.....

Date of Submission : 11 May 2026
Date of Defense : 3 June 2026

To my family,

FOREWORD

I would like to express my gratitude to Assoc. Prof. Dr. Ahmet Onat for his guidance, mentoring and constant assistance.

I also would like to thank to Begüm Seçkin and Talha Enes Ercan for their support during whole education process.

Finally, I would like to express my appreciation to executives and colleagues from AVL Research and Engineering Türkiye for their support and flexibility.

June 2026

Emir OĞUZ
(Senior Systems Engineer)

TABLE OF CONTENTS

Page

FOREWORD ..... **ix**
TABLE OF CONTENTS ..... **xi**
ABBREVIATIONS ..... **xiii**
SYMBOLS ..... **xv**
LIST OF TABLES ..... **xvii**
LIST OF FIGURES ..... **xix**
SUMMARY ..... **xxi**
ÖZET ..... **xxv**
1. INTRODUCTION **1**
1.1 Motivation of Thesis 2
1.2 Purpose of Thesis 3
1.3 Literature Review 5
2. FUEL CELL SYSTEM AND MATHEMATICAL MODELLING **9**
2.1 Fuel Cell System Overview 9
2.1.1 Electrochemical reaction principles 10
2.1.2 Fuel cell losses 12
2.1.3 PEMFC system architecture 13
2.1.4 Fuel cell system control concepts 15
2.1.5 Application areas 15
2.2 FC System Modelling 16
2.2.1 Cathode flow model 17
2.2.2 Compressor dynamics 17
2.2.3 System power 18
2.2.4 Oxygen excess ratio definition 18
3. REINFORCEMENT LEARNING AND DEEP Q NETWORK **21**
3.1 Markov Decision Process 21
3.2 Value Function and Bellman Equation 22
3.3 Deep Q Network 25
4. SIMULATION STUDIES AND RESULTS **29**
4.1 Control Architectures 29
4.1.1 Direct Deep Q Network 30
4.1.2 PI-augmented Deep Q Network control 31
4.2 Training 33
4.2.1 Learning environment parameters 33
4.2.2 Reward design 35
4.3 Testing 38
5. CONCLUSIONS **51**
REFERENCES **55**
CURRICULUM VITAE **59**

ABBREVIATIONS

DC	**:** Direct Current
DRL	**:** Deep Reinforcement Learning
DQN	**:** Deep Q-Network
EILMMA-DDPG	**:** Enhanced Independent Learning Multi-Agent Deep Deterministic Policy Gradient
FC	**:** Fuel Cell
LADRC	**:** Linear Active Disturbance Rejection Control
LQC	**:** Linear Quadratic Qubic
LQR	**:** Linear Quadratic Regulator
MAS-FQL	**:** Multi-Agent Fuzzy Q-Learning
MDP	**:** Markov Decision Process
MESD-DDPG	**:** Multi-role Exploration Strategy Distributed Deep Deterministic Policy Gradient
MRAC	**:** Model Reference Adaptive Control
OER	**:** Oxygen Excess Ratio
PEMFC	**:** Proton Exchange Membrane Fuel Cell
PI	**:** Proportional Integral
RL	**:** Reinforcement Learning
ULM	**:** Ultra-Local Model
SO	**:** Snake Optimization
SMC	**:** Sliding Mode Control
TD	**:** Temporal Difference

SYMBOLS

$\mathbf{A}$	**:** Action space in MDP
$\mathbf{a}$	**:** Action
$\mathbf{a'}$	**:** Next action
$\mathbf{E_0}$	**:** Reversable cell voltage
$\mathbf{E_N}$	**:** Nerst cell voltage
$\mathbf{E_r}$	**:** Reversable cell voltage
$\mathbf{F}$	**:** Faraday constant
$\mathbf{G_t}$	**:** Return from time t
$\mathbf{h}$	**:** Ratio of specific heat capacities of air
$\mathbf{I_i}$	**:** Current of i
$\mathbf{i}$	**:** Current density
$\mathbf{i_0}$	**:** Exchange current density
$\mathbf{i_L}$	**:** Limiting current
$\mathbf{K_t}$	**:** Torque constant
$\mathbf{l}$	**:** Length
$\mathbf{m_i}$	**:** Mass of i
$\mathbf{n}$	**:** Moving number of electrons
$\mathbf{P}$	**:** State transition probability function
$\mathbf{P_i}$	**:** Power of i
$\mathbf{p_i}$	**:** Pressure of i
$\mathbf{Q}$	**:** Energy
$\mathbf{Q^{\pi}}$	**:** Action value function under policy π
$\mathbf{R}$	**:** Resistance
$\mathbf{R_{const}}$	**:** Gas constant
$\mathbf{r}$	**:** Reward
$\mathbf{S}$	**:** State space in MDP
$\mathbf{s}$	**:** Current state
$\mathbf{s_t}$	**:** Observation state
$\mathbf{s'}$	**:** Next state

$\mathbf{T}$	**:** Temperature
$\mathbf{t}$	**:** Time
$\mathbf{t_s}$	**:** Sampling time
$\mathbf{U_i}$	**:** Voltage of i
$\mathbf{V_i}$	**:** Volume of i
$\mathbf{V^{\pi}}$	**:** State value function under policy π
$\mathbf{w_i}$	**:** Mass flow rate of i
$\mathbf{W_i}$	**:** Reward weight of i
$\mathbf{y}$	**:** Target value
$\mathbf{\cancel{A}}$	**:** Area
$\boldsymbol{\alpha}$	**:** Learning rate
$\boldsymbol{\alpha_{coef}}$	**:** Charge transfer coefficient
$\boldsymbol{\gamma}$	**:** Discount factor
$\mathbf{\Delta S}$	**:** Entropy change
$\mathbf{\Delta Q}$	**:** Energy change
$\mathbf{\Delta G}$	**:** Gibbs free energy change
$\mathbf{\Delta H}$	**:** Enthalpy change
$\boldsymbol{\varepsilon}$	**:** Exploration rate
$\boldsymbol{\eta_{act}}$	**:** Activation loss
$\boldsymbol{\eta_{conc}}$	**:** Concenration loss
$\boldsymbol{\eta_{ohm}}$	**:** Ohmic loss
$\boldsymbol{\theta}$	: Temporal difference
$\boldsymbol{\lambda}$	**:** Oxygen Excess Ratio
$\boldsymbol{\pi}$	**:** Policy
$\boldsymbol{\rho}$	**:** Resistivity coefficient
$\boldsymbol{\tau}$	**:** Target smooth factor
$\boldsymbol{\tau_{load}}$	**:** Compressor torque
$\boldsymbol{\omega}$	**:** Compressor speed

LIST OF TABLES

Page

Table 4.1 : PI controller configurations. **29**
Table 4.2 : Training parameters. **34**
Table 4.3 : Reward weights. **37**
Table 4.4 : Performance comparison. **48**
Table 4.5 : Timing performance comparison. **49**

LIST OF FIGURES

Page

Figure 2.1 : Structure of a PEMFC [8]. **10**
Figure 2.2 : PEMFC architecture [9]. **14**
Figure 2.3 : Maximum power points [11]. **19**
Figure 2.4 : Relation between stack current and OER [11]. **19**
Figure 3.1 : Agent environment interaction [30]. **21**
Figure 3.2 : Neural network [40]. **25**
Figure 4.1 : PI controller structure. **29**
Figure 4.2 : Direct DQN control structure. **30**
Figure 4.3 : Direct DQN non-linear action space. **31**
Figure 4.4 : PI augmented DQN control structure. **32**
Figure 4.5 : Decay of randomness in exploration exploitation trade-off (ε_t). **33**
Figure 4.6 : High steady state OER error without integral term. **38**
Figure 4.7 : Oscillatory compressor current. **39**
Figure 4.8 : Training rewards. **40**
Figure 4.9 : Training Q values. **41**
Figure 4.10 : Compressor current. **41**
Figure 4.11 : Measured OER. **42**
Figure 4.12 : OER error. **42**
Figure 4.13 : Net power output. **43**
Figure 4.14 : Measured OER with classical controllers. **43**
Figure 4.15 : Training reward with mistuned PI controllers. **45**
Figure 4.16 : Training Q values with mistuned PI controllers. **46**
Figure 4.17 : Measured OER with mistuned PI controllers. **47**

FUEL CELL AIR SUPPLY SYSTEM CONTROL USING DEEP Q-NETWORK

SUMMARY

The increasing global population and economic growth are causing a continuous rise in energy consumption. Energy production remains an indispensable component of industrial development for every country. However, traditional fossil fuel sources such as coal, oil, and natural gas are causing serious environmental problems due to their rapid depletion as well as greenhouse gas emissions that accelerate global warming. Considering the deteriorating environmental conditions, rising global temperatures, and greenhouse gas emissions, the transition from fossil fuels to renewable energy sources has become imperative.

In this context, clean and efficient energy sources such as solar, wind, geothermal, and hydrogen emerge as the most viable alternatives to replace conventional sources. Particularly hydrogen exhibits great potential as both a fuel and an energy carrier and storage medium due to the intermittent nature of solar and wind energy. Through the carbon-free alternatives it offers, hydrogen is increasingly being recognized as a globally viable energy source and fuel alternative.

Hydrogen is positioned as the primary fuel for a new generation power generation device that produces electricity, heat, water, and zero pollution using oxidant and hydrogen. The fuel cell (FC) operates through an electrochemical redox reaction: the reaction of hydrogen and oxygen produces energy, water, and heat. At the anode, hydrogen undergoes catalytic decomposition to form protons and electrons; while protons pass through the proton exchange membrane to the cathode, electrons are transferred to the external circuit to provide energy to the power load. At the cathode, oxygen reacts with protons and electrons to produce water and heat.

When evaluating the environmental impacts of mobility vehicles, three main propulsion systems emerge: internal combustion engines, battery electric vehicles, and FC electric vehicles. Internal combustion engines produce significant amounts of carbon dioxide, nitrogen oxides, and particulate matter emissions as they are based on the direct combustion of fossil fuels. Battery electric vehicles, while offering zero tailpipe emissions, have disadvantages such as limited range and long charging times.

Proton exchange membrane fuel cells (PEMFC) have become a prominent technology, particularly in the automotive sector, due to advantages such as zero emissions, low noise, fast response time, long range, and short refueling time. PEMFC systems are electrochemical devices that directly convert chemical energy into electrical energy and find a wide range of applications from portable power applications to hybrid electric vehicles and FC electric vehicles, with characteristics of high efficiency, high power density, low operating temperature, and minimal pollution.

PEMFC systems generally consist of 5 subsystems: thermal management subsystem, water management system, hydrogen supply subsystem, air supply subsystem, and power electronics management subsystem. While the hydrogen supply subsystem uses fast electrical valves to minimize the pressure difference between the anode and cathode, the air supply subsystem provides the necessary oxygen to the cathode through a compressor. Since water and temperature dynamics change much more slowly compared to air flow dynamics, many studies assume that the FC operates under constant temperature conditions. The power electronics subsystem controls load connection and includes application-specific converters.

The most critical component of the air supply system is the air compressor. For stable operation of the FC, the presence of sufficient oxygen at the cathode side is mandatory. The oxygen excess ratio (OER, denoted with λ) is defined as the ratio of the amount of oxygen supplied to the cathode to the amount of oxygen consumed in the electrochemical reaction and is one of the fundamental indicators of system efficiency.

Due to the slow air flow dynamics of the compressor and air supply manifolds, when the load current changes suddenly, oxygen starvation can occur at the cathode if sufficient oxygen cannot be rapidly supplied. This situation can lead to the formation of hot spots on the membrane surface and irreversible damage to polymeric membranes. Additionally, oxygen starvation causes serious problems such as low power generation, membrane degradation, and even shortening of stack operating life.

The manifold system is modeled as a collective volume representing all pipes that supply gas to the stack and drain gas from the stack. The supply manifold pressure depends on downstream cathode dynamics, and the dynamics of cathode variables depend on both the air from the supply manifold and the amount of external current load. Based on the ideal gas assumption, the partial pressures of oxygen and nitrogen can be derived through differential equations.

OER control is of great importance for both the economic efficiency and safety of PEMFC systems. The system faces challenges such as nonlinearity, parametric uncertainty, and load disturbances. During rapid load current increase, the chemical reaction within the FC accelerates substantially, and this situation can lead to excessive oxygen consumption and even oxygen starvation that can cause permanent cell damage.

The OER at the cathode must be kept greater than 1; however, this should not be interpreted to mean that higher OER always implies better performance. Conversely, excessively high OER causes the air compressor to consume more power and consequently reduces net power output. Net power is defined as the difference between stack power generation and parasitic power consumed by the compressor, and at high loads, these parasitic losses can reach up to thirty percent of the FC's power consumption.

Research shows that maximum net power for different stack currents and OER values is obtained when OER is between 2 and 2.5. Therefore, the primary objective of OER control is to prevent oxygen starvation while minimizing compressor power consumption and thus optimizing system efficiency. The optimal OER reference varies depending on load current and is generally found in a value range between 1.5 and 3.

One of the greatest challenges encountered in control design of PEMFC systems is the complex and highly nonlinear dynamic structure of the system. Traditional model-based control approaches require an accurate system model; however, PEMFCs exhibit nonlinear and time-varying uncertainties due to aging and degradation. This situation creates significant limitations as the performance of methods such as model predictive control (MPC) is largely dependent on an accurate system model.

In this context, reinforcement learning-based approaches offer significant advantages. The Q-learning algorithm is a model-free approach that determines an optimal policy without requiring estimation of the environment's dynamics. This method works in a data-driven manner to address challenges such as system nonlinearity, parametric uncertainty, and load disturbances, and does not require characterization of system dynamics.

The findings demonstrate that PI and reinforcement learning collaboration markedly improves transient OER control by increasing response speed and diminishing both safety-critical OER violations and efficiency-reducing deviations, especially when the baseline PI controller is inadequately calibrated. The decline in steady-state accuracy in specific configurations underscores the necessity for precise reward shaping and action constraint design to prevent ongoing corrective behavior approaching equilibrium. The findings indicate that hybrid control performance is fundamentally dependent on configuration and support that reinforcement learning is most effective as a supplementary corrective mechanism rather than a standalone alternative for classical PI control in PEMFC air management systems, where durability and efficiency must be concurrently maintained.

DERİN Q-AĞI KULLANILARAK YAKIT HÜCRESİ HAVA BESLEME SİSTEMİ KONTROLÜ

ÖZET

Dünya genelinde artan nüfus ve ekonomik büyüme, enerji tüketiminin sürekli olarak artmasına neden olmaktadır. Enerji üretimi, her ülkenin endüstriyel kalkınmasının vazgeçilmez bir bileşeni olmaya devam etmektedir. Ancak kömür, petrol ve doğal gaz gibi geleneksel fosil yakıt kaynakları, hızla tükenmelerinin yanı sıra küresel ısınmayı hızlandıran sera gazı emisyonları nedeniyle ciddi çevresel sorunlara yol açmaktadır. Kötüleşen çevre koşulları, yükselen küresel sıcaklıklar ve sera gazı emisyonları göz önüne alındığında, fosil yakıtlardan uzaklaşarak yenilenebilir enerji kaynaklarına geçiş zorunlu hale gelmiştir.

Bu bağlamda güneş, rüzgâr, jeotermal ve hidrojen gibi temiz ve verimli enerji kaynakları, konvansiyonel kaynakların yerine geçebilecek en uygulanabilir alternatifler olarak öne çıkmaktadır. Özellikle hidrojen, güneş ve rüzgâr enerjisinin kesintili doğası nedeniyle hem yakıt hem de enerji taşıyıcısı ve depolama aracı olarak büyük potansiyel sergilemektedir. Sunduğu karbonsuz alternatifler sayesinde hidrojen, küresel ölçekte uygulanabilir bir enerji kaynağı ve yakıt alternatifi olarak giderek daha fazla kabul görmektedir.

Hidrojen, oksidant ve hidrojen kullanılarak elektrik, ısı, su ve sıfır kirlilik üreten yeni nesil bir güç üretim cihazının temel yakıtı olarak konumlanmaktadır. Yakıt hücresi, elektrokimyasal bir redoks reaksiyonu aracılığıyla çalışır: hidrojen ve oksijenin reaksiyonu sonucunda enerji, su ve ısı üretilir. Anotta hidrojen katalitik ayrışmaya uğrayarak protonlar ve elektronlar oluşturur; protonlar proton değişim membranı üzerinden katoda geçerken, elektronlar dış devreye aktarılarak güç yüküne enerji sağlar. Katotta ise oksijen, protonlar ve elektronlarla reaksiyona girerek su ve ısı üretir.

Mobilite araçlarının çevresel etkileri değerlendirildiğinde, üç temel tahrik sistemi karşımıza çıkmaktadır: içten yanmalı motorlar, bataryalı elektrikli araçlar ve yakıt hücreli elektrikli araçlar. İçten yanmalı motorlar, fosil yakıtların doğrudan yanmasına dayandığından önemli miktarda karbondioksit, azot oksitler ve partikül madde emisyonu üretmektedir. Bataryalı elektrikli araçlar ise sıfır egzoz emisyonu sunmakla birlikte, sınırlı menzil ve uzun şarj süreleri gibi dezavantajlara sahiptir.

Proton değişim membran yakıt hücreleri (PEMFC), sıfır emisyon, düşük gürültü, hızlı tepki süresi, uzun menzil ve kısa yakıt ikmali süresi gibi avantajları nedeniyle özellikle otomotiv sektöründe öne çıkan bir teknoloji haline gelmiştir. PEMFC sistemleri,

kimyasal enerjiyi doğrudan elektrik enerjisine dönüştüren elektrokimyasal cihazlar olup, yüksek verimlilik, yüksek güç yoğunluğu, düşük çalışma sıcaklığı ve minimum kirlilik özellikleriyle taşınabilir güç uygulamalarından hibrit elektrikli araçlara ve yakıt hücreli elektrikli araçlara kadar geniş bir uygulama alanı bulmaktadır.

PEMFC sistemleri genel olarak alt sistemden oluşmaktadır: sıcaklık yönetimi alt sistemi, su yönetim sistemi, hidrojen besleme alt sistemi, hava besleme alt sistemi ve güç elektroniği yönetimi alt sistemi. Hidrojen besleme alt sistemi, anot ve katot arasındaki basınç farkını minimize etmek için hızlı elektriksel valfler kullanırken, hava besleme alt sistemi kompresör aracılığıyla katoda gerekli oksijeni sağlar. Su ve sıcaklık dinamikleri hava akış dinamiklerine kıyasla çok daha yavaş değiştiğinden, birçok çalışmada yakıt hücresinin sabit sıcaklık koşullarında çalıştığı varsayılmaktadır. Güç elektroniği alt sistemi yük bağlanmasını kontrol eder ve uygulamaya özgü dönüştürücüler içerir.

Hava besleme sisteminin en kritik bileşeni hava kompresörüdür. Yakıt hücresinin kararlı çalışması için katot tarafında yeterli oksijenin bulunması zorunludur. Oksijen fazlalık oranı (OER, λ ile ifade edilir), katoda sağlanan oksijen miktarının elektrokimyasal reaksiyonda tüketilen oksijen miktarına oranı olarak tanımlanır ve sistem verimliliğinin temel göstergelerinden biridir.

Kompresör ve hava besleme manifoldlarının yavaş hava akış dinamikleri nedeniyle, yük akımı ani değiştiğinde, yeterli oksijen hızlı bir şekilde sağlanamazsa katotta oksijen açlığı meydana gelebilir. Bu durum, membran yüzeyinde sıcak noktaların oluşmasına ve polimerik membranlarda geri dönüşü olmayan hasara yol açabilir. Ayrıca oksijen açlığı, düşük güç üretimi, membran bozulması ve hatta yığın çalışma ömrünün kısalması gibi ciddi sorunlara neden olmaktadır.

Manifold sistemi, yığına gaz sağlayan ve yığından gaz boşaltan tüm boruları temsil eden toplu bir hacim olarak modellenmektedir. Besleme manifoldu basıncı, akış aşağı katot dinamiklerine bağlıdır ve katot değişkenlerinin dinamikleri hem besleme manifoldundan gelen havaya hem de harici akım yükü miktarına bağlıdır. İdeal gaz varsayımına dayalı olarak, oksijen ve azotun kısmi basınçları diferansiyel denklemler aracılığıyla türetilebilir.

OER kontrolü, PEMFC sistemlerinin hem ekonomik verimliliği hem de güvenliği açısından büyük önem taşımaktadır. Sistem doğrusal olmama, parametrik belirsizlik ve yük bozuklukları gibi zorluklarla karşı karşıyadır. Hızlı yük akımı artışı anında, yakıt hücresi içindeki kimyasal reaksiyon büyük ölçüde hızlanır ve bu durum aşırı oksijen tüketimine, hatta kalıcı hücre hasarına neden olabilecek oksijen açlığına yol açabilir.

Katottaki OER'nin 1'den büyük tutulması gerekmektedir; ancak bu, daha yüksek OER'nin her zaman daha iyi performans anlamına geldiği şeklinde yorumlanmamalıdır. Aksine, aşırı yüksek OER, hava kompresörünün daha fazla güç tüketmesine ve dolayısıyla net güç çıktısının azalmasına neden olur. Net güç, yığın

güç üretimi ile kompresör tarafından tüketilen parazitik güç arasındaki fark olarak tanımlanır ve yüksek yüklerde bu parazitik kayıplar yakıt hücresinin güç tüketiminin yüzde otuzuna kadar ulaşabilir.

Araştırmalar, farklı yığın akımları ve OER değerleri için maksimum net gücün, OER 2 ile 2.5 arasındayken elde edildiğini göstermektedir. Bu nedenle OER kontrolünün temel amacı, oksijen açlığını önlerken kompresör güç tüketimini minimize etmek ve böylece sistem verimliliğini optimize etmektir. Optimal OER referansı, yük akımına bağlı olarak değişmekte olup genellikle 1.5 ile 3 arasında bir değer aralığında bulunmaktadır.

PEMFC sistemlerinin kontrol tasarımında karşılaşılan en büyük zorluklardan biri, sistemin karmaşık ve yüksek derecede doğrusal olmayan dinamik yapısıdır. Geleneksel model tabanlı kontrol yaklaşımları, doğru bir sistem modeli gerektirmektedir; ancak PEMFC'ler, yaşlanma ve bozulma nedeniyle doğrusal olmayan ve zamanla değişen belirsizlikler sergilemektedir. Bu durum, model öngörücü kontrol (MPC) gibi yöntemlerin performansının büyük ölçüde doğru sistem modeline bağlı olması nedeniyle önemli sınırlamalar yaratmaktadır.

Bu bağlamda, pekiştirmeli öğrenme (reinforcement learning) tabanlı yaklaşımlar önemli avantajlar sunmaktadır. Q-öğrenme algoritması, ortamın dinamiklerinin tahminini gerektirmeden optimal bir politika belirleyen model-bağımsız bir yaklaşımdır. Bu yöntem, sistemin doğrusal olmama, parametrik belirsizlik ve yük bozuklukları gibi zorluklarını ele almak için veri odaklı bir şekilde çalışır ve sistem dinamiklerinin karakterizasyonuna ihtiyaç duymaz.

Elde edilen bulgular, özellikle temel PI denetleyicinin yetersiz kalibre edildiği durumlarda, PI ve pekiştirmeli öğrenme iş birliğinin geçici rejimde OER kontrol performansını belirgin biçimde iyileştirdiğini göstermektedir. Bu iyileşme, yanıt hızının artmasıyla birlikte hem güvenlik açısından kritik OER ihlallerinin hem de verim kaybına yol açan sapmaların anlamlı ölçüde azalması şeklinde ortaya çıkmaktadır. Bununla birlikte, bazı yapılandırmalarda kararlı hâl doğruluğunda gözlenen düşüş, denge noktasına yaklaşırken devam eden düzeltici davranışların önüne geçilebilmesi için ödül fonksiyonunun ve eylem kısıtlarının titizlikle tasarlanması gerektiğine işaret etmektedir. Sonuçlar, hibrit kontrol mimarisinin başarımının büyük ölçüde yapılandırmaya bağlı olduğunu ve pekiştirmeli öğrenmenin, dayanıklılık ve verimliliğin eş zamanlı olarak sağlanmasının kritik olduğu PEMFC hava yönetim sistemlerinde, klasik PI denetiminin yerine geçen bağımsız bir çözümden ziyade tamamlayıcı bir düzeltme mekanizması olarak daha etkili olduğunu ortaya koymaktadır.

1. INTRODUCTION

The increasing global population and economic growth are causing a continuous rise in energy consumption. Energy production remains an indispensable component of industrial development for every country. However, traditional fossil fuel sources such as coal, oil, and natural gas are causing serious environmental problems due to their rapid depletion as well as greenhouse gas emissions that accelerate global warming. Considering global warming, carbon emission quantity must be reduced. Thus, an environment friendly and sustainable energy resource must be understood more and clearly such as hydrogen energy and Fuel cells (FC) [2]. The International Energy Agency forecasts a 30% increase in worldwide energy demand by 2040, with transportation sector accounting for approximately 25% of global CO2 emissions.

In this context, clean and efficient energy sources such as solar, wind, geothermal, and hydrogen emerge as the most viable alternatives to replace conventional sources. Particularly hydrogen exhibits great potential as both a fuel and an energy carrier and storage instrument due to the discontinuous nature of solar and wind energy. Through the carbon-free alternatives it offers, hydrogen is increasingly being recognized as a globally viable energy source and fuel alternative.

When evaluating the environmental impacts of mobility vehicles, three main propulsion systems emerge: internal combustion engines, battery electric vehicles, and fuel cell electric vehicles. Internal combustion engines produce significant amounts of carbon dioxide, nitrogen oxides, and particulate matter emissions as they are based on the direct combustion of fossil fuels. Battery electric vehicles, while offering zero tailpipe emissions, have disadvantages such as limited range and long charging times.

Thanks to its advantageous characteristics—including high efficiency, high power density, relatively low operating temperature, silent operation, and near-zero emissions—a hydrogen-fueled proton exchange membrane fuel cell (PEMFC) is widely accepted for a broad range of applications [10]. Despite these advantages,

several technical challenges must be addressed before PEMFCs can achieve widespread commercial adoption.

1.1 Motivation of Thesis

While PEMFCs present significant advantages for automotive applications, their transition from laboratory to commercial deployment faces a critical barrier which is the durability challenge, which is directly related to the long-term durability and reliability of the system. Degradation occurring in the FC stack accelerates particularly during operation under variable load conditions and significantly shortens the system lifespan. Degradation mechanisms such as chemical and mechanical deterioration in the polymeric membrane structure, catalytic activity loss, and carbon corrosion directly affect the system performance.

The oxygen excess ratio (OER, denoted with λ) is defined as the ratio of the amount of oxygen supplied to the cathode to the amount of oxygen consumed in the electrochemical reaction and is considered one of the fundamental indicators of system performance. Keeping the OER value greater than one is mandatory to prevent oxygen starvation; however, when this ratio rises to excessively high levels, the parasitic power consumed by the air compressor increases significantly. Operating below OER=1.5 increases degradation rate by 40-60%, while exceeding OER=3.0 results in 25-30% parasitic power loss, particularly under high load conditions, and thus the net power output of the system decreases seriously [27].

One of the most significant triggers of aging in PEMFC systems is the oxygen starvation condition experienced at the cathode. When the load current increases suddenly, if sufficient oxygen cannot reach the cathode in time due to the slow dynamics of the air supply system and this situation leads to irreversible polymeric membrane damage. Oxygen starvation also accelerates carbon support corrosion, causes catalytic activity to decrease, and ultimately results in premature failure of the FC stack. Therefore, maintaining the OER at an appropriate reference value is crucial to system durability and reliability. To mitigate oxygen starvation and maintain optimal performance, precise control of oxygen supply is essential [11].

Net power is calculated as the difference between the gross power produced by the stack and the parasitic power consumed by the compressor and other auxiliary

equipment. Studies conducted in the literature show that maximum net power is achivied when the OER value is between 2 and 2.5 [28]. This situation reveals that OER control must be optimized not only to prevent aging but also to maximize system efficiency. Therefore, OER control requires a delicate balance between compressor power consumption and oxygen starvation risk. Achieving and maintaining this optimal OER range under dynamic load conditions, however, requires sophisticated control strategies considering multiobjective control problem.

Historically, the PI controller has been the standard approach for OER regulation. In this approach, the deviation of OER from the reference value is calculated to produce an error signal, and this error signal is processed by the PI controller to generate the compressor motor current command. However, this method contains several significant disadvantages. First, the gains of the PI controller must be adjusted manually, and these adjustments are generally performed through trial-and-error methods. Manual PI tuning requires quite a few hours of expert time and results in suboptimal performance [7]. Secondly, due to the highly nonlinear dynamic structure and time-varying parametric uncertainties of PEMFC systems, fixed-gain PI controllers cannot exhibit optimal performance under all operating conditions.

Based on these motivations, the development of a model-free and adaptive control algorithm that directly produces the compressor motor current command is of great importance in terms of both increasing system reliability. Reinforcement learning (RL) based approaches offer a suitable solution for this purpose due to their ability to learn optimal control policies through interaction with the environment without requiring complete knowledge of the system model. Building upon the identified challenges and the potential of RL approaches, this thesis establishes the improvement of transient response.

1.2 Purpose of Thesis

The fundamental objective of this thesis study is to develop an RL-based, model-free control algorithm for the optimal regulation of the OER in PEMFC systems. The proposed approach directly generates the compressor motor current command from the OER reference by implicitly accounting for the operating conditions corresponding to maximum net power output, thereby eliminating the conventional PI controller layer. Rather than reducing algorithmic complexity, this formulation shortens the

control decision chain and removes intermediate tuning-dependent structures, which improves responsiveness and facilitates real-time implementation. Thereafter, the hybrid control structure of PI and RL-based combination is investigated.

This thesis aims to establish of a simulation environment representing the dynamic behavior of the PEMFC system and the training of an RL algorithm on this environment. This simulation environment will be designed to reflect the realistic behavior of the system under variable load conditions and will serve as the training platform for RL agents.

It will be investigated the applicability of modern RL methods such as Q-learning and Deep Q-Networks (DQN) algorithms for PEMFC OER control and comparatively evaluating the performances of these algorithms. Particularly, the capabilities of the algorithms to maintain OER at the reference value against rapid load changes and their performances in minimizing compressor power consumption considering aging will be examined.

One of the aims of this thesis is to compare the developed RL based controller with the traditional PI controller in the literature and numerically demonstrating its superiorities. This comparison will be conducted in terms of OER tracking error, net power generation, and prevention of oxygen starvation events criteria.

The main contributions of this thesis include the systematic design of a reinforcement learning environment, including specific state representations, action space formulation, and a reward structure that explicitly balances OER regulation, system efficiency, and operational safety. Furthermore, a hybrid control architecture combining a classical PI controller with a DQN agent is proposed, enabling improved transient performance and robustness against load variations and model uncertainties. The research offers a comprehensive comparative analysis of pure learning-based and hybrid control strategies, illustrating that the proposed method improves response speed, decreases OER deviation, and minimizes risks such as starvation. Finally, the thesis contributes to the literature by showing the practical applicability of model-free deep reinforcement learning methods in complex, nonlinear energy systems where accurate modeling is challenging, thereby offering a scalable and adaptive alternative to traditional model-based control techniques.

As a result of this study, the development of a control strategy that both increases the durability of PEMFC systems and optimizes energy efficiency, which can be easily implemented in industrial applications, is expected.

1.3 Literature Review

One of the main technical challenges that limits the wider adoption of FCs as a reliable power source is their durability under diverse operating conditions. Depending on the application, the expected service life of an FC can differ considerably—around 5,000 hours for automotive use, up to 20,000 hours for buses, and as much as 40,000 hours of continuous operation in stationary systems. [37]. When the oxygen flow at the cathode is insufficient, localized hot spots can develop on the polymer membrane. These hot spots not only diminish cell performance by restricting reactant availability at the triple-phase boundary but also accelerate membrane degradation, ultimately shortening the overall lifetime of the FC. Although performance decay in fuel cells cannot be completely avoided, its progression can be significantly slowed by developing a thorough understanding of the underlying degradation processes and failure mechanisms [37]. When the load current increases, oxygen consumption in the electrochemical reactions rises accordingly, which can result in oxygen starvation.

In PEMFCs, the air supply system—driven by a compressor—responds relatively slowly due to its inherent dynamics. In addition, this subsystem requires a considerable amount of electrical energy, which reduces the overall net power output of the FC. [35]. The control of the air supply system is designed to sustain the optimal OER while preventing oxygen starvation, particularly during sudden changes in load current. [1].

By regulating the air flow to the cathode, the supply of reactant oxygen can be balanced to optimize both the stack power output and the compressor's energy consumption. In addition, controlling the oxygen flow provides a means to mitigate the risks of oxygen starvation and membrane dehydration [6]. It is well established that the OER achieves its maximum efficiency at the optimal value of minimum OER = 2 [28].

The net power of an FC system P_{net} is the difference between the power generated by the stack P_{st} and the parasitic power consumed by auxiliary components. In automotive fuel cells, most of this parasitic power is used by the air compressor, making the proper regulation of air flow a key factor in maximizing overall efficiency. Increasing the

OER raises oxygen partial pressure, which improves both stack power P_{st} and net power P_{net}. However, once the optimum level is reached, any further increase leads to only small gains in stack power while the compressor consumes much more energy. As a result, the net power output decreases. A static controller is insufficient for mitigating the negative effects of disturbances, such as variations in stack load current. [28]. From an energy management perspective, the air supply system can consume nearly 25% of the power produced by the stack. In order to minimize these losses and maximize net output, it is essential to adjust the OER within an appropriate scale [41].

Conventional PI controllers have been widely employed for OER regulation due to their implementation of simplicity and well-understood tuning methodologies. However, PI controllers with fixed gains cannot achieve optimal performance across the entire operating envelope due to the nonlinear nature of the plant. Manual PI tuning requires quite a few hours of expert time and results in suboptimal performance [7]. Furthermore, the PI controller structure introduces intermediate computational steps between the OER error signal and the compressor motor command, increasing control algorithm complexity.

Gaining a clear understanding of the inherently nonlinear behavior of the FC is essential for accurate modeling and effective control design. [34]. Designing LQG controllers for the system requires linearizing the model around a specific operating point, which increases computational load. The LQR control strategy relies on access to the full state vector; however, in practice, not all system states can be directly measured [26].

The analytical model of the linearized dynamics retains certain structural characteristics of the physical system. However, once numerical values are substituted to represent a specific equilibrium point, these structural properties are lost. As a result, the validity of the analysis is confined to the vicinity of that operating point. [34].

In [16], a feedback linearization approach was applied to regulate the air supply system under the assumption of no system uncertainties. [4]. To account for system uncertainties, [31] proposed an adaptive controller incorporating a state observer.

When there is power fluctuations in demand, classical control applications will diminish the controller's robustness [21]. Also, PI controllers are not able achieve optimal control with nonlinear systems. Considering model-free approach in RL

algorithms, although the system is nonlinear the agent can be feed with multi observation state and multi action can be expected to obtain optimal actions depending on current state of the system while sustaining global search ability [16]. However, the training process might require longer steps depending on the complexity of the system [25].

Studies in [20] and [19] investigated adaptive fuzzy output-feedback control strategies for uncertain nonlinear systems subject to state constraints. Studies in [18] focuses on deep reinforcement learning (DRL) control strategies such as Deep Deterministic Policy Gradient for the thermal management to regulate stack temperature and Proximal Policy Optimization for the air supply system to control cathode pressure and OER. The work in [22] shows that sliding mode control (SMC), and model reference adaptive control (MRAC) ensures robust OER control with reduced control input including fault-tolerant features. The study in [17] presents a DRL-based intelligent control framework for air and hydrogen supply. The proposed approach is based on the Enhanced Independent Learning Multi-Agent Deep Deterministic Policy Gradient (EILMMA-DDPG) algorithm and is shown to mitigate Q-value overestimation while satisfying real-time control requirements under varying operating conditions.

In [39], an ultra-local model (ULM) control structure combined with multi-agent system fuzzy Q-learning (MAS-FQL) is proposed to dynamically regulate the oxygen excess ratio. The results demonstrate that the ULM-based MAS-FQL approach achieves effective OER management under different operating scenarios.

In [5] authors propose a fusion control strategy combining Linear Active Disturbance Rejection Control (LADRC), Proportional Integral Differential (PID) control, and the Snake Optimization (SO) algorithm to regulate OER and they conclude proposed method has better overshoot and settling time.

The study in [15] introduces a distributed DRL based adaptive PI controller for air flow control in PEMFCs. A Multi-role Exploration Strategy Distributed Deep Deterministic Policy Gradient (MESD-DDPG), is proposed to tune the PI controller coefficients. The study shows that tuning is needed for this kind of controller considering distinctive conditions in PEMFCs.

2. FUEL CELL SYSTEM AND MATHEMATICAL MODELLING

2.1 Fuel Cell System Overview

The FCs are devices that convert chemical energy to electrical energy without any intermediate step. There are no moving parts for the operation, and they work quietly. FCs are able to work on lower temperatures. Unlike conventional heat engines, which are thermodynamically constrained by the Carnot cycle efficiency limit, FCs operate through direct energy conversion pathways, enabling potentially higher conversion efficiencies. When supplied with hydrogen as the primary fuel, the only by-products of the electrochemical reaction are water and heat, thereby positioning FCs as zero-emission power generation systems with significant environmental advantages.

FC operation was discovered by a Swiss scientist in 1830s, and it was possible obtain low amount of electrical power [8]. Technology is enhanced year by year and the first PEMFC which is the most favorite type of FC was built during 1960s.

Among various FC technologies, PEMFCs have emerged as particularly suitable candidates for automotive propulsion applications due to their favorable operational characteristics [28]. These characteristics include high power density, solid-state electrolyte configuration, extended operational lifespan, minimal corrosion issues, and relatively low operating temperature range of 50 to 100 degrees Celsius. The low operating temperature eliminates the necessity for extensive thermal insulation while enabling safer operation.

The fundamental structure of a PEMFC comprises a polymer electrolyte membrane positioned between two porous electrodes designated as the anode and cathode as illustrated in Figure 2.1. Also, they have an electrolyte which carries the ions from one electrode to other one, and a catalyst to increase the speed of reaction [32].

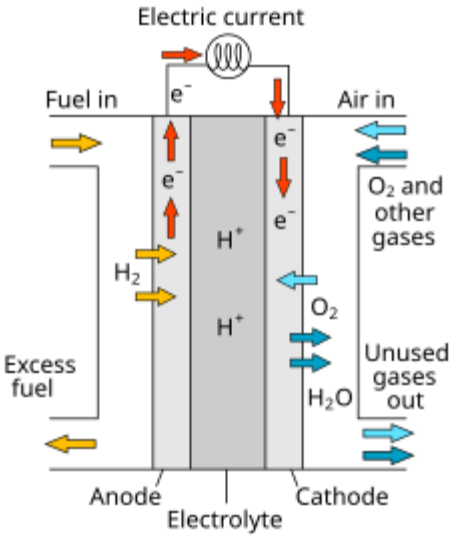

Figure 2.1 : Structure of a PEMFC [8].

One cell can only produce a small amount of voltage, but they can be connected to series and this assembly is named as a stack. Stack design can be configured according to application; cell number can be adjusted and can be grouped parallel.

FCs and batteries seem similar in structural perspective but indeed there are several points that they differ. First and the most obvious difference is the way of providing electrical energy. While batteries storage devices, FCs generate electrical current continuously. Although batteries can be refilled, an external circuit is needed for this operation which means operation disruption. On the other hand, fuel cells can continue producing electrical energy as fuel and oxidant are fed into the system. Secondly, the energy density of fuel cells is much higher than batteries, which makes them suitable for long-term applications.

Since FCs transform fuel directly into electrical energy, less waste heat is produced, which contributes to high efficiency. Moreover, produced heat energy can be used for household, water or industrial heating.

2.1.1 Electrochemical reaction principles

The electrochemical reactions occurring within a PEMFC constitute the fundamental mechanism for electrical power generation. In general, gas fuel goes to the anode side and oxidant goes to the cathode side. While electrons complete the external circuit, hydrogen and oxygen are used to make electricity, along with water as a by-product.

At the anode electrode, hydrogen molecules undergo catalytic oxidation, dissociating into protons and electrons according to the half-reaction expressed as in equation 2.1.

$$H_2 \rightarrow 2H^+ + 2e^- \tag{2.1}$$

The protons that are made move through the polymer electrolyte membrane to the cathode electrode. The electrons, on the other hand, move through an external electrical circuit, doing useful work by creating electrical current. An oxygen reduction reaction happens at the cathode when oxygen molecules from the air stream mix with protons and electrons that are moving. The end result is water as in equation 2.2.

$$4H^+ + 4e^- + O_2 \rightarrow 2H_2O \tag{2.2}$$

The main redox reaction in an FC is expressed in equation 2.3, the combination of hydrogen and oxygen to make water, electricity, and heat:

$$2H_2 + O_2 \rightarrow 2H_2O \tag{2.3}$$

A single hydrogen-oxygen FC has a theoretical open-circuit voltage of about 1.23 volts when thermodynamic conditions are perfect. But when the system is running with electrical loads connected, the terminal voltage drops a lot, to a range of 0.6 to 0.7 volts, because of different irreversible losses [12].

Enthalpy represents the portion of energy unavailable for useful work. For an exothermic reaction, enthalpy is equal to energy released in the reaction. For an endothermic reaction, it is the energy absorbed from the surroundings and expressed with ΔH.

Entropy is the amount of energy that excessed from the reaction other than useful work. In FC world, this is mostly heat energy that does not contribute to the reaction and expressed as in equation 2.4.

$$\Delta S = \frac{\Delta Q}{T} \tag{2.4}$$

Gibbs free energy is the useful work in a thermodynamical system. The Gibbs free energy is the amount of the work that is exchanged with the its environment and can be expressed with equation 2.5.

$$\Delta G = G_f\, of\ prodcuts - G_f of\ reactants = \Delta H - T\Delta S \tag{2.5}$$

In an FC, voltage of a cell is directly related with Gibbs free energy with equation 2.6. The reversible cell potential, also referred to as the Gibbs potential under standard conditions, represents the maximum achievable open-circuit voltage of the fuel cell.

$$\Delta G = -nFE_r \tag{2.6}$$

Where, n is the moving number of electrons, F is Faraday constant, E_{rev} is ideal cell voltage.

Nernst potential is the open circuit voltage that can be achieved by an FC in an ideal case. Measured cell voltage cannot go up to this voltage level because of the losses and is calculated as in equation 2.7. The Nernst potential represents the reversible cell voltage under non-standard operating conditions, accounting for temperature and partial pressure effects.

$$E_N = E_0 - \frac{R_{const}T}{nF} lnQ \tag{2.7}$$

2.1.2 Fuel cell losses

The deviation between the theoretical reversible cell voltage and the actual operating voltage arises from three principal categories of losses: activation losses, ohmic losses, and concentration losses.

Activation loss is required over potential to start the reaction by crossing the activation energy barrier. Its effect is more perceivable the FCs that operates in low temperature. In high temperature FCs most likely to not bother from this scope. This type of loss also affects the speed of the reactions. Moreover, its proportion is the biggest one among loss types. The relationship between activation overvoltage and current density is commonly described by the Tafel equation as shown in equation 2.8 [13].

$$\eta_{act} = \frac{R_{const}T}{\alpha_{coef}nF} \ln\left(\frac{i}{i_0}\right) \tag{2.8}$$

Ohmic loss is related with the resistance of electrolyte, electrode, and current collectors during the transportation of electrons. The voltage drop associated with ohmic losses exhibits a linear relationship with current density [32], as demonstrated in equaitons 2.9 and 2.10. This loss can be decreased with better geometry.

$$\eta_{ohm} = iR \tag{2.9}$$

$$R = \frac{\rho l}{\text{A}} \tag{2.10}$$

Concentration losses, also termed mass transport losses, is caused by the hydrogen fraction where the reactions occur. When the concentration decreases, it causes insufficient transportation of reactant to the electrode. It is critical at high current demands. These losses become particularly pronounced at high current densities and can lead to severe performance deterioration if adequate reactant supply is not maintained, as expressed in equation 2.11 [28].

$$\eta_{conc} = \frac{R_{const}T}{nF} \ln\left(1 - \frac{i}{i_L}\right) \tag{2.11}$$

2.1.3 PEMFC system architecture

A complete PEMFC system encompasses the FC stack as the central power generation unit along with multiple auxiliary subsystems as shown in Figure 2.2, essential for maintaining proper operating conditions. These subsystems include the air supply system, hydrogen supply system, thermal management system, water management system, and power electronics system [28]. The coordinated operation of these subsystems is critical for achieving efficient, reliable, and safe FC system performance.

Heat management is one of the key subjects in FC systems. It includes both a cooling and a heating system. FCs are able to start in cold environments with the help of electrical heaters. At the beginning of the freeze start-up operation electricity is provided to electrical heaters. At this point of the operation current is not drawn from the stack. When the temperature rises to a suitable level for the operation, then hydrogen is supplied and electrical power is generated. On the other hand, during the operation heat is produced and temperature can go beyond operation

conditions. To avoid this situation, cooling fans and water pumps are used to reduce the temperature. It also affects the stack temperature, inlet, and outlet gas temperatures.

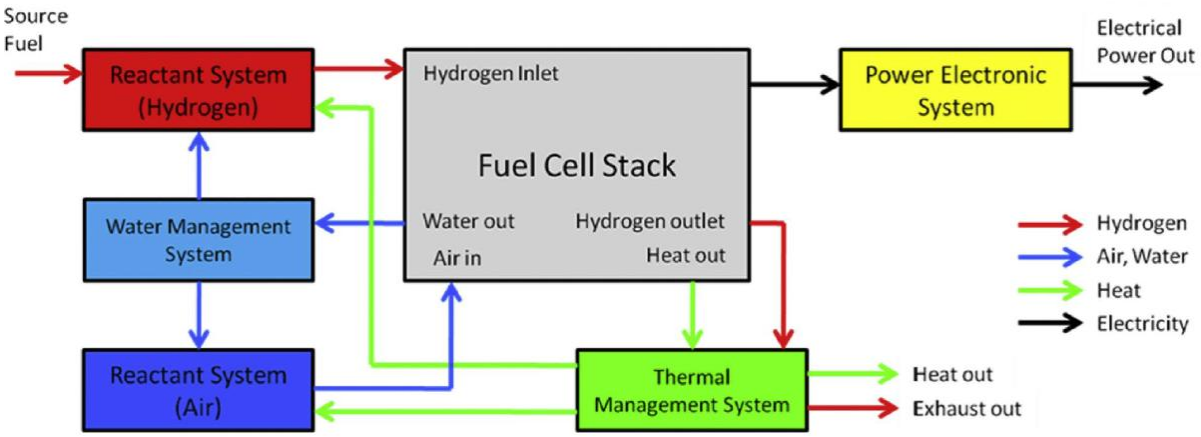

Figure 2.2 : PEMFC architecture [9].

The last product of the reaction, which is water, needs to be drained from the system in a safe way to secure the operation. Even though water is produced on the cathode side of the cell, it is possible to water molecules may diffuse to anode side depending on concentration difference. Another functionality of the path is to adjust the humidity of the inlet gases to the stack. It is an important parameter because insufficient humidity can damage the membrane.

Hydorgen supply subsystem handles hydrogen supply to the system coming from a hydrogen tank via controlling valves and compressors in the path. With changing current demand hydrogen flow must be adapted using valve commands. The path also consists hydrogen recirculation blowers which are used to bring in the unused hydrogen to the system instead of exhausting them.

Similar to the hydrogen supply system, the air supply path is responsible for supplying air to the system by controlling the compressor. If there is not enough oxygen in the system, it causes the FC to degrade. On the flipside running, the compressor on a high level consumes too much power and reduces net power output, which means decreasing efficiency. That is why OER is a crucial parameter to control, and this thesis focuses on its control.

The air supply system dynamics are characterized by a relatively slow response compared to the fast electrochemical processes within the FC stack. The time constants associated with compressor inertia and manifold filling dynamics are on the order of 10^0 to 10^1 seconds, whereas

electrochemical reactions occur at time scales of 10^{-19} seconds [28]. This temporal mismatch creates significant challenges for maintaining adequate oxygen supply during rapid load transients, which forms the primary motivation for advanced control strategy development.

The power control system evaluates the stack current depending on power demand calculation for the actuators done according to this value. Also, connecting and cutting the load to the system is controlled by this subsystem via contactor or relays. An application specific converter can be placed against the need of the modification of electrical power.

2.1.4 Fuel cell system control concepts

The start-up phase of a PEMFC system is a very active and dangerous time, when the air, hydrogen, humidification, and thermal subsystems all need to be controlled at the same time. In steady-state operation, system variables change slowly. In contrast, start-up involves quick changes and a lot of uncertainty about the parameters.

For reliable operation, fault prevention, and degradation mitigation, it is important to constantly check the states of the system. A strict monitoring system usually has the following important parts: OER, cathode and anode pressure and differential pressure, stack voltage, stack current, stack power, cell voltage, humidification and membrane water content, temperature distribution, and the status of the compressor and actuator.

Shutdown is another critical transient phase that significantly influences long-term durability. Repeated improper shutdown cycles are documented to cause catalyst dissolution, carbon corrosion, and membrane thinning, especially under air/hydrogen boundary formation.

Safety constraints shape every aspect of PEMFC control, from operating limits to fault handling strategies. Key considerations include oxygen starvation prevention, avoiding parasitic losses and over-oxidation, managing pressure differential limits, hydrogen leak detection, actuator saturation and windup protection, fault-tolerant control and redundancy, avoiding going beyond critical temperature regions.

2.1.5 Application areas

In FC systems, particularly those utilized in automotive and mobile applications, managing transient load demands poses a considerable challenge owing to the sluggish dynamic response of FCs relative to traditional energy sources. Combining FCs with supercapacitors, which can

deliver high power density for short periods of time, is one way to solve this problem. By absorbing or delivering peak currents, supercapacitors can protect the FC stack from quick changes and slow down performance degradation. This architecture not only makes the system more responsive, but it also extends the life of the FC by preventing oxygen starvation and reducing voltage undershoots. [38] assert that the integration of supercapacitors in PEMFC systems has resulted in enhanced dynamic load tracking and increased overall system efficiency by alleviating the immediate demand on the FC stack.

Considering their operation conditions PEMFCs are the most suitable type of FC to be used in automotive industry. For commercial use, different types of brands have their FC electric vehicles: Honda Clarity, Toyota Mirai, Hyundai ix35 FCEV, and the Hyundai Nexo. It is also possible and more suitable to implement PEMFCs to buses, trucks and forklifts taking into account their long-distance usage.

To be used for primal power source and backup unit in hospitals, universities, data centres and large office buildings, there are stationary applications which is a growing sector day by day.

FC applications are available in different types of marine vessels such as cruise ships, hybrid ferries and submarines. Limited storage space problem can be overcome with water electrolysis [33].

World's biggest plane manufacturers have also put some effort on FCs to use them in aerial applications. Boeing and its European partners carried out a flight powered by an FC combining with batteries on a propeller plane. Airbus has a programme called ZEROe which aims to manufacture a propeller plane powered by FC. Clean Aviation is also one of the biggest programmes co-funded by the European Union, which objects to decrease net emission [8].

2.2 FC System Modelling

Modelling is one of the necessities to develop a control algorithm. It allows us to understand the process and predict the behavior of the related system under different operational conditions. It is a great way to test without using real hardware, which is not always possible due to high costs. Obviously, a certain match cannot be expected between model and real system but still it is widely used and beneficial.

2.2.1 Cathode flow model

In order to model cathode flow dynamics, oxygen, nitrogen, water mass flows can be defined with the equations 2.12, 2.13 and 2.14 [28]:

$$\frac{d}{dt}m_{O_2} = w_{O_2,in} - w_{O_2,out} - w_{O_2,react} \tag{2.12}$$

$$\frac{d}{dt}m_{N_2} = w_{N_2,in} - w_{N_2,out} \tag{2.13}$$

$$\frac{d}{dt}m_{wtr,cathd} = w_{vpr,cathd,in} - w_{vpr,cathd,out} + w_{vpr,genr} + w_{vpr,memb} \tag{2.14}$$

The supply manifold and return manifold pressure dynamics can be calculated with the help of the equations 2.15 and 2.16 [28]:

$$\frac{d}{dt}p_{sm} = \frac{hR}{M_a V_{sm}}(w_{comp}T_{comp} - w_{sm}T_{sm}) \tag{2.15}$$

$$\frac{d}{dt}p_{rm} = \frac{RT_{fc}}{M_a V_{rm}}(w_{cath} - w_{rm}) \tag{2.16}$$

2.2.2 Compressor dynamics

The compressor constitutes the primary actuator for air supply regulation, and its dynamic behavior fundamentally determines the system response characteristics. The compressor is typically driven by a brushless DC motor whose rotational speed governs the air mass flow rate delivered to the FC stack. The relationship between motor input and compressor output is governed by electromechanical dynamics that can be expressed through the torque balance equation as in equation 2.17 [28]:

$$J\frac{d\omega}{dt} + B\omega = K_t i - \tau_{load} \tag{2.17}$$

where J represents the combined compressor and motor inertia, ω denotes the rotational speed, B is the viscous friction coefficient, K_t is the motor torque constant, i represents the motor current, and τ_{load} corresponds to the compressor load torque. In current-controlled configurations, the motor current serves as the direct control input, providing more predictable dynamic response characteristics compared to voltage-controlled approaches.

The compressor power to pump air can be defined as in equation 2.18:

$$P_{comp} = I_{comp} U_{comp} \tag{2.18}$$

2.2.3 System power

Considering the portion of the compressor power, minimal losses can be neglected and net power of the system can be expressed with equation 2.19.

$$P_{net} = P_{st} - P_{comp} \tag{2.19}$$

where P_{st} is the stack electrical power output and P_{comp} represents the compressor power consumption. In automotive PEMFC systems, the compressor constitutes the primary source of parasitic power, potentially consuming up to 25 to 30 percent of the stack output under high load conditions [27].

2.2.4 Oxygen excess ratio definition

The OER, serves as a critical performance indicator characterizing the balance between oxygen supply and consumption within the FC cathode, as calculated with equation 2.20. The OER is defined as the ratio of the oxygen mass flow rate supplied to the cathode to the oxygen consumption rate determined by the electrochemical reaction stoichiometry [28]:

$$\lambda = \frac{w_{O_2,in}}{w_{O_2,react}} \tag{2.20}$$

where $W_{O_2,\mathrm{in}}$ represents the oxygen mass flow rate entering the cathode and $W_{O_2,\mathrm{react}}$ denotes the oxygen consumption rate by the electrochemical reaction.

Research has quantified the relationship between OER levels and degradation rates. Operating below OER values of 1.5 has been shown to increase degradation rates by 40 to 60 percent

compared to operation at optimal conditions. Conversely, maintaining excessively high OER values above 3.0 results in parasitic power losses of 25 to 30 percent while providing negligible benefits to stack performance [11]. These findings underscore the critical importance of precise OER control within the optimal operating window.

It is a known fact that there is a relationship between net power output and OER, as can be seen in Figure 2.3. In order to use PEMFC in the most efficient way, the OER must be adjusted depending on the power demand. Especially at higher loads maintaining lower OER compared to lower loads makes the system more efficient [28].

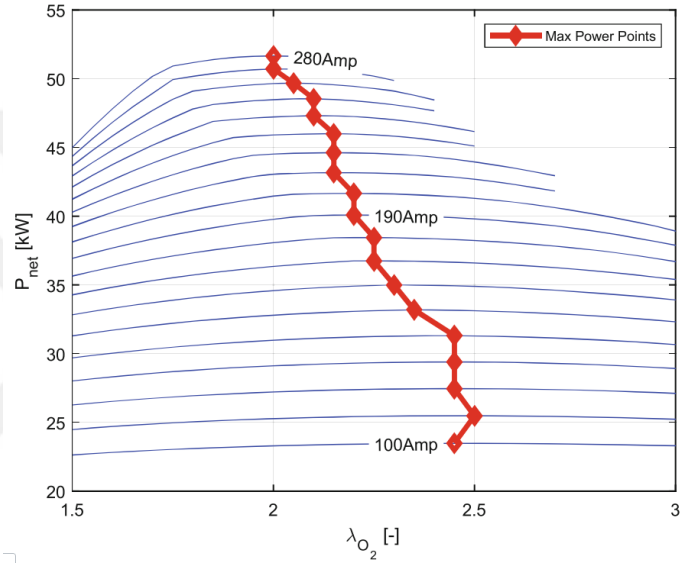

Figure 2.3 : Maximum power points [11].

Analogously, a relationship between stack current and OER can be established as in Figure 2.4 to be used for generating optimal OER values during operation.

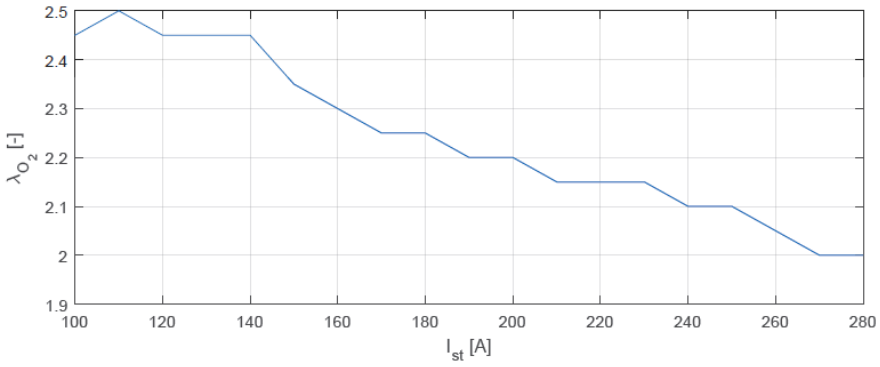

Figure 2.4 : Relation between stack current and OER [11].

3. REINFORCEMENT LEARNING AND DEEP Q NETWORK

This section presents the conceptual and mathematical foundations of RL. RL is a learning concept in which an autonomous agent interacts with its environment and acquires optimal decision-making strategies through trial-and-error guided by rewards and penalties. RL is gaining popularity day by day because it does not require an explicit model of system dynamics; it has become a powerful alternative to classical control strategies, especially nonlinear and time-varying systems [30].

3.1 Markov Decision Process

In RL, learning is formulated as a dynamic interaction between an agent and an environment as shown in Figure 3.1. At each discrete time step, the agent receives an observation of the current state S, selects an action A, and the environment responds by generating a reward R and transitioning the system into a new state. This sequential decision-making process is formally described by a Markov Decision Process (MDP).

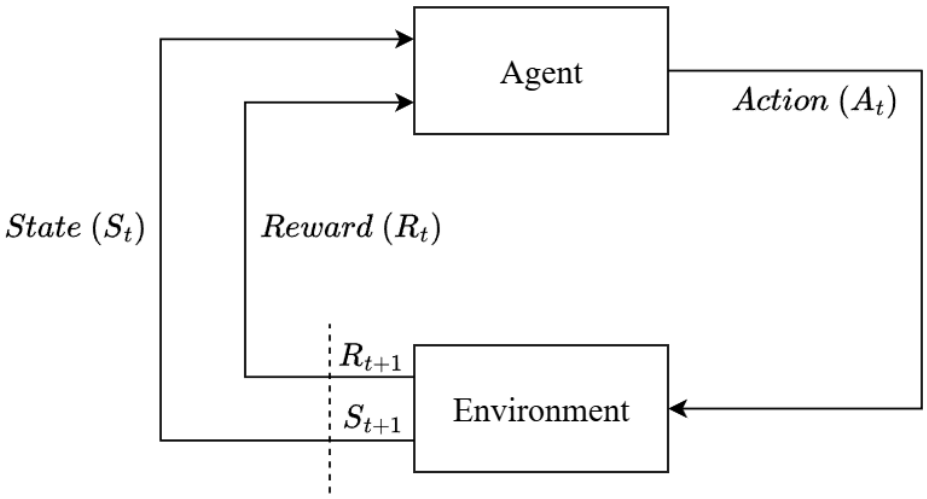

Figure 3.1 : Agent environment interaction [30].

An MDP provides the mathematical framework upon which RL is formally defined. An MDP models sequential decision-making problems where subsequent states rely solely on the present state and action, independent of the prior sequence of events. This characteristic is known as the Markov property [3]. Formally, an MDP is represented as the tuple $\mathrm{M} = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R})$ where S: Set of all possible states, A: Set of available actions, $\mathrm{P}(\mathrm{s}'|\mathrm{s}, \mathrm{a})$: Transition probability describing the likelihood of moving to state s' after taking action a at state s, $\mathrm{R}(\mathrm{s}, \mathrm{a})$: Reward function that provides a scalar feedback signal [29].

The agent interacts with the MDP by choosing actions according to some policy. A sequence of interactions generates a trajectory: $\mathrm{s}_0, \mathrm{a}_0, \mathrm{r}_0, \mathrm{s}_1, \mathrm{a}_1, \mathrm{r}_1, \ldots$ The Markov property allows the development of recursive equations, such as the Bellman equations, which are essential to most RL algorithms. This assumption underlies the theoretical guarantees of most RL algorithms and allows the development of recursive equations, such as the Bellman equations, which are essential to most RL algorithms [30].

The states in the MDP framework for PEMFC OER control are constructed depending on system dynamics. For instance, the physical quantities such as pressure, mass flow, fraction of a particular molecule, power, etc. can be combined and defined as states. The action decided by the agent is the control signal which is compressor current. The transition dynamics are constituted by FC dynamics, and these dynamics are implemented in a programming and numerical computing platform.

3.2 Value Function and Bellman Equation

The reward function is the key mechanism guiding the agent toward desirable behavior. It quantitatively encodes the objective of the learning task. A well-designed reward encourages stable and efficient learning, whereas a poorly defined reward may lead to suboptimal, unstable, or unsafe policies [24]. In control tasks, reward structures typically incorporate with minimizing tracking error, reducing energy consumption, ensuring the system remains within safe operational bounds, limiting overshoot, undershoot, or oscillations, penalizing aggressive or unsafe actions.

The aim of the agent is to maximize collected summed reward (return) during its lifetime and this return is defined as the discounted sum of rewards at a certain time t is calculated with equation 3.1.

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3.1}$$

where γ adjusts the relative importance of future rewards relative to the present [36].

A policy $\pi(s) = a$ defines the behavior of an agent depending on the current state of the system, which is defined in equation 3.2. It determines how actions are selected based on the current state. The Q-value calculation described in equation 3.3, or action-value function, quantifies the expected return when taking action a in state s and following a policy π thereafter.

$$V^{\pi}(s) = \mathbb{E}\left[G_t \mid s_0 = s\right] \tag{3.2}$$

$$Q^{\pi}(s,a) = \mathbb{E}\left[G_t \mid s_0 = s, a_0 = a\right] \tag{3.3}$$

Using the law of total expectation and the Markov property, and with the help of Bellman expectation equation [3] considering its recursive relationship with these value functions, the state-value function under policy π depending on the probability next state s' and gained reward for a certain state R and selected action a is expressed with equation 3.4.

$$V^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V^{\pi}(s')] \tag{3.4}$$

The Bellman optimality equations are satisfied by the optimal value functions $V^*(s)$ and $Q^*(s,a)$ defined for the optimal policy π^*. Equation 3.5 shows the equation that applies directly to the action-value function. Through its one-step forward effects, this expression demonstrates how the maximum operator can characterize the optimal action in a given state [34].

$$Q^*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} Q^*(s',a')\right] \tag{3.5}$$

The Bellman equation describes the value of a state depending on a determined policy as the discounted summation of the expected reward. These recursive equations form the theoretical backbone of algorithms such as Q-learning, DQN, and policy iteration.

A defining difficulty in RL is balancing exploration and exploitation. Choosing only exploratory or exploitative behavior leads to suboptimal learning. With probability ε, a random action is selected to explore, and with probability 1−ε the action with the highest Q-value is selected to exploit. Typically, ε is gradually reduced to ensure early exploration and late-stage convergence. In constant ε-greedy strategy ε remains fixed through training. In decay ε-greedy ε decreases over time exponentially as calculated in equation 3.6.

$$\varepsilon_t = \varepsilon_0 e^{-kt} \tag{3.6}$$

In linear decay ε decreases over time linearly as shown in equation 3.7.

$$\varepsilon_{t+1} = \max\left(\varepsilon_{min}, \varepsilon_t - \Delta\varepsilon\right) \tag{3.7}$$

In adaptive ε-greedy ε is adjusted based on learning stability or uncertainty measures [36].

Experience replay stores past interactions $(\mathrm{s}, \mathrm{a}, \mathrm{r}, \mathrm{s}')$ in a replay buffer and samples from batches of experiences for training. This mechanism provides several benefits such as breaking correlation between consecutive samples, improving sample efficiency, and stabilizing training with function approximation.

Training and testing phases are clearly separated in RL based control applications. During the training phase, the agent interacts with the environment, updates the policy, and learns from rewards. During the test phase, the policy is fixed, and system performance is evaluated. Typical evaluation metrics include tracking accuracy, stability in transient and steady-state regimes, overshoot and undershoot magnitudes, energy efficiency, compliance with safety and operational constraints. So, a thorough testing stage is essential for validating RL controllers in real-world applications.

Tabular RL becomes impractical in high-dimensional environments. Therefore, modern RL methods approximate value functions and policies using neural networks, leading to DRL. The DRL is able to work with nonlinear, complex behavior, across large state spaces systems and in compliance with continuous control tasks [14].

3.3 Deep Q Network

A table of Q-values is kept for each state-action pair in standard Q-learning. But this method doesn't work well for state spaces that are big or always changing. [23] came up with DQN to get around this problem. DQN uses deep neural networks to get close to the Q-value function. The state s is what goes into the network, and the Q-values for all possible actions are what comes out. This architecture enables DQN to generalize across extensive state spaces and proficiently capture intricate environmental features.

The neural network is constructed by a combination of perceptrons shown in Figure 3.2 with input, hidden, and output layers. Number of input layers is associated with the number of observation states that agent takes, and output layer is associated with the number of actions that agent performs. On the other hand, hidden layers can be constructed depending on the implementation. It is possible to change the number of layers and perceptrons. Increasing the numbers can lead to obtaining more accurate results, but this makes operation more complex, which means more computational load. Nevertheless, having less perceptrons and layers might not be enough for complicated problems.

Through backpropagation and gradient-based optimization, the neural network learns a compact and generalizable approximation of the action-value function. This capability allows DQN to operate effectively in large-scale and continuous state spaces where traditional tabular methods are infeasible.

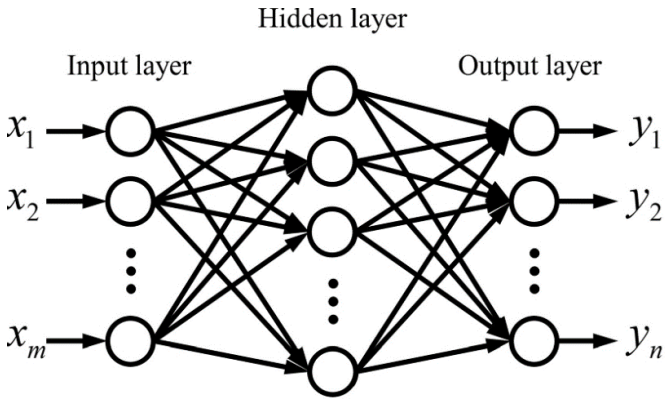

Figure 3.2 : Neural network [40].

Since Q^* is unknown, the DQN minimizes the discrepancy between the current network prediction and a target value derived from the Bellman equation. For a given transition s_t, a_t, r_t, s_{t+1} the target Q-value is defined as equation 3.8 [23].

$$y_t = r_t + \gamma \max_{a'} Q\,(s_{t+1}, a'; \theta) \tag{3.8}$$

The Bellman equation is used by the DQN algorithm to update Q-values over and over again. The network, on the other hand, does not store Q-values directly. Instead, it reduces the temporal difference (TD) error, which is the difference between the predicted Q-value and the target Q-value by updating its weights. The neural network parameters are optimized by minimizing the mean squared error between the predicted Q-value and the target is calculated with equation 3.9.

$$Loss = (y_t - Q(s_t, a_t; \theta))^2 \tag{3.9}$$

Although DQN has demonstrated strong performance in high-dimensional reinforcement learning problems, they suffer from a systematic overestimation bias in action-value estimates. This issue arises from the maximization operation in the Bellman target, where the same network is used both to select and evaluate actions. As a result, estimation errors are amplified, leading to overly optimistic Q-values and potentially unstable learning behavior.

To address this limitation, the Double DQN algorithm was introduced as an extension of DQN. The key idea behind Double DQN is to decouple action selection from action evaluation during target value computation, thereby reducing overestimation bias while preserving the overall structure of value-based learning.

In standard DQN, the target Q-value is computed according to equation (3.8). Here, the same network parameters θ are used both to determine the maximizing action and estimate its corresponding value.

Here, the same network parameters θ are used to both determine the maximizing action and estimate its corresponding value.

Double DQN mitigates this issue by introducing two neural networks: an online network with parameters θ and a target network with parameters θ^-. The online network is responsible for selecting the best action, while the target network evaluates the value of that selected action.

The target Q-value in Double DQN is defined with equation 3.11.

$$y_t = r_t + \gamma Q(s_{t+1}, arg\max_{a'} Q(s_{t+1}, a'; \theta); \theta^-) \tag{3.10}$$

Similar to standard DQN, Double DQN minimizes the mean squared TD error between the predicted Q-value and the target value. Double DQN is particularly beneficial in environments with large action spaces, where Q-value overestimation can significantly degrade performance. From an implementation perspective, Double DQN introduces minimal additional computational complexity compared to standard DQN, as it only requires maintaining a separate target network. Consequently, it has become a widely adopted baseline in modern DRL applications, including control systems, robotics, and energy management problems.

In control-oriented RL tasks, such as FC air supply or oxygen excess ratio regulation, overestimated Q-values may result in aggressive or oscillatory control actions. The more conservative and accurate value estimates provided by Double DQN contribute to smoother control behavior and improved system robustness, making it a suitable choice for safety-critical and energy-sensitive applications.

4. SIMULATION STUDIES AND RESULTS

In this thesis simulation and test studies are held on MATLAB and Simulink environment. Two different architectures are trained and then testing of these architectures are completed.

4.1 Control Architectures

This research examines two separate DQN-based control frameworks for the regulation of PEMFC OER. The initial method utilizes the RL agent as an independent controller that directly produces the compressor motor current directive. The second architecture combines the DQN agent with a traditional PI controller in a parallel arrangement, wherein the agent generates corrective modifications to enhance the fundamental PI control signal.

During training with hybrid structure and testing different PI controller configurations are used as shown in Table 4.1 and PI controller is implemented as in Figure 4.1. Some of the configurations are miscalibrated intentionally to comparison.

Table 4.1 : PI controller configurations.

Configuration	K_p	K_i
Configuration 1 – Balanced	30	345
Configuration 2 – Robust	12	80
Configuration 3 – Aggressive	121	1812
Configuration 4 – Mistuned robust	15	690
Configuration 5 – Mistuned aggressive	60	172.5

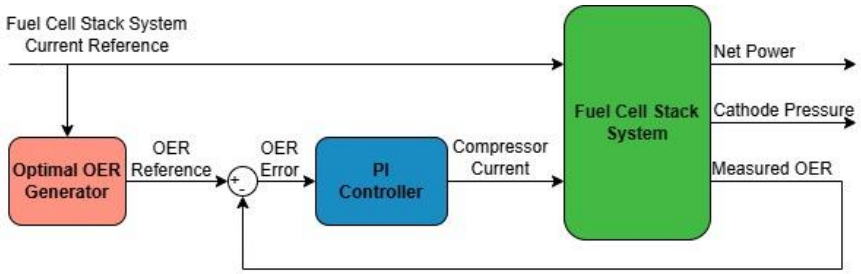

Figure 4.1 : PI controller structure.

4.1.1 Direct Deep Q Network

The agent takes 9 observation states within the direct DQN control framework, which include measured OER, OER error, integral of OER error, derivative of OER error, cathode pressure, derivative of cathode pressure, reference OER, agent action (compressor current), and settling information, as in equation 4.1. Control structure of direct DQN control is shown in Figure 4.2.

$$s_t = \begin{bmatrix} \lambda_{measured}(t) \\ \lambda_{err}(t) \\ \int_0^t \lambda_{err}(\tau)d\tau \\ \frac{d}{dt}\lambda_{err}(t) \\ p_{cath}(t) \\ \frac{d}{dt}p_{cath}(t) \\ \lambda_{ref}(t) \\ I_{comp}(t) \\ \sigma_{settle}(t) \end{bmatrix} \tag{4.1}$$

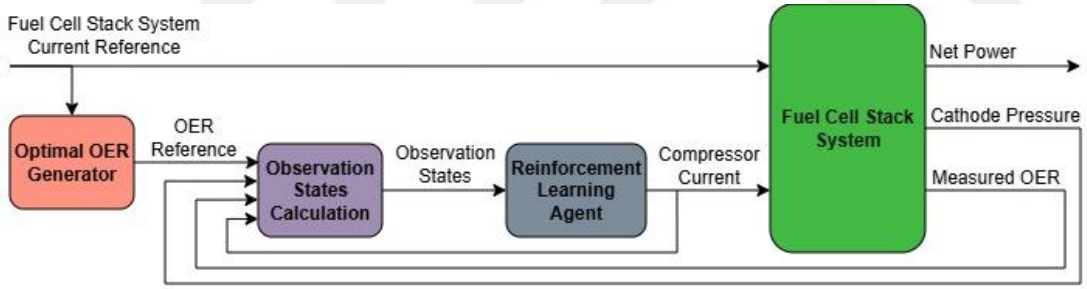

Figure 4.2 : Direct DQN control structure.

One of the main contributions of this thesis lies in the structured design of the action space through a non-uniform discretization strategy that explicitly incorporates prior knowledge of the system's operating characteristics. Rather than identifying the operating regions themselves, which are already well-established for the fuel cell air supply system, this study utilizes these known nominal operating points to guide the reinforcement learning process. Specifically, the compressor current domain [0,62] A is partitioned around two empirically and physically significant nominal operating points, p1 = 19.435 A and p2 = 42.11 A, which correspond to typical load regimes of the system.

This targeted discretization results in a finite set of 29 distinct control actions, where a higher concentration of action candidates is allocated in the vicinity of these critical operating regions, while a coarser representation is maintained in less relevant zones. Such a design enables the learning agent to achieve finer control granularity where sensitivity to actuation is high, thereby improving policy precision during normal operation, without unnecessarily increasing the dimensionality of the action space.

From a control perspective, this approach enhances both learning efficiency and control performance: it accelerates convergence by focusing exploration on practically relevant regions and reduces oscillatory or suboptimal behavior caused by insufficient resolution near steady-state conditions. As illustrated in Figure 4.3, the resulting action distribution reflects a deliberate bias toward nominal operating points, effectively embedding domain knowledge into the reinforcement learning framework. This contribution demonstrates that informed action space shaping can significantly improve the applicability of deep reinforcement learning in nonlinear, safety-critical energy systems.

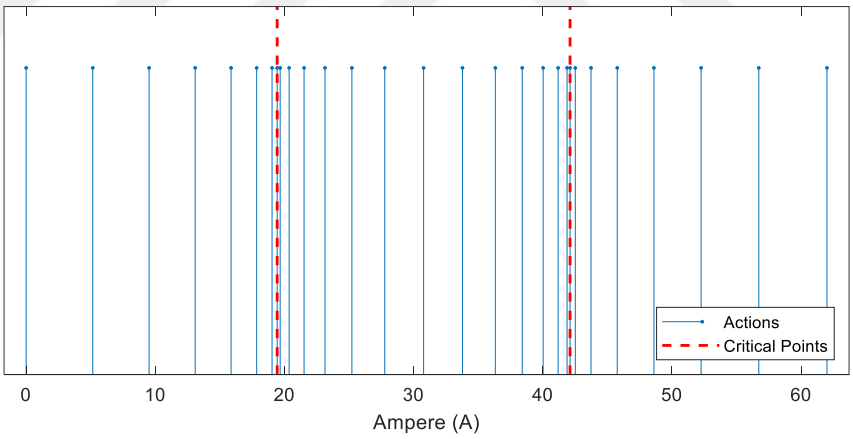

Figure 4.3 : Direct DQN non-linear action space.

4.1.2 PI-augmented Deep Q Network control

The hybrid architecture integrates a traditional PI controller with the DQN agent in a parallel manner. The PI controller delivers core tracking performance, but the RL agent produces corrective modifications. The composite control signal is demonstrated with equation 4.2.

$$I_{comp} = I_{PI} + \Delta u_{RL} \tag{4.2}$$

where I_{PI} indicates the output of the PI controller and Δu_{RL} indicates the corrective action of the RL agent, as control architecture is illustrated in Figure 4.4. Compared to the direct DQN control, the observation space is enhanced by including the PI controller output and the compressor current, resulting in an 11-dimensional state vector as in equation 4.3.

$$s_t = \begin{bmatrix} \lambda_{measured}(t) \\ \lambda_{err}(t) \\ \int_0^t \lambda_{err}(\tau)d\tau \\ \frac{d}{dt}\lambda_{err}(t) \\ p_{cath}(t) \\ \frac{d}{dt}p_{cath}(t) \\ \lambda_{ref}(t) \\ I_{comp}(t) \\ \sigma_{settle}(t) \\ I_{PI}(t) \\ \Delta u_{RL} \end{bmatrix} \tag{4.3}$$

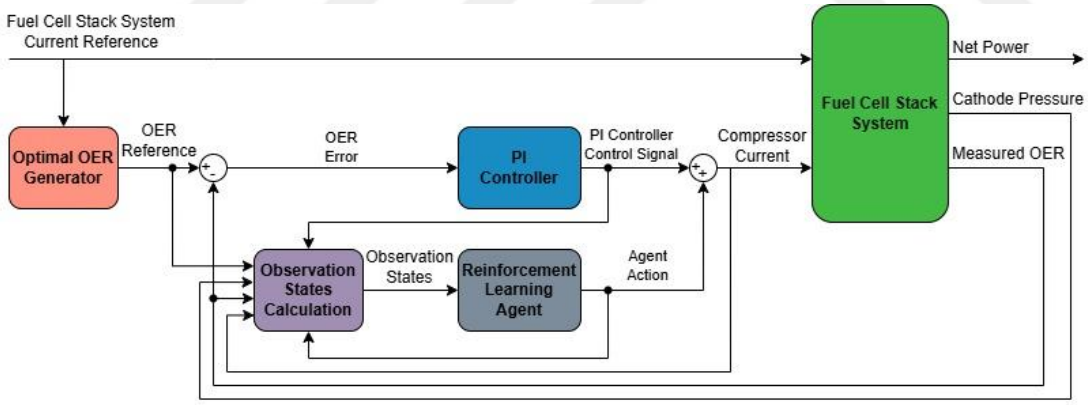

Figure 4.4 : PI augmented DQN control structure.

The action space for the corrective signal consists of 25 discrete values divided into three categories: fine adjustments $\Delta u \in \{0,\pm1,\pm3,\pm5\}$A for steady-state refinement, moderate corrections $\Delta u \in \{\pm10, \pm12.5, \pm15, \pm17.5, \pm20\}$A for transient compensation, and aggressive interventions $\Delta u \in \{\pm30, \pm40, \pm50, \pm62\}$A for rapid disturbance rejection.

4.2 Training

The critic network architecture comprises three fully linked hidden layers of 64, 64, and 32 neurons. The dimensions of the output layer align with the dimension of the discrete action space.

The exploration rate is set to ε_t = 1.0 in the beginning, and an exponential decay schedule descends ε to the minimum value ε_{min} = 0.08 over 80% of the total training steps, as shown in Figure 4.5. This plan enables the agent to gather varied experiences during initial training while gradually leveraging acquired information in the subsequent phases.

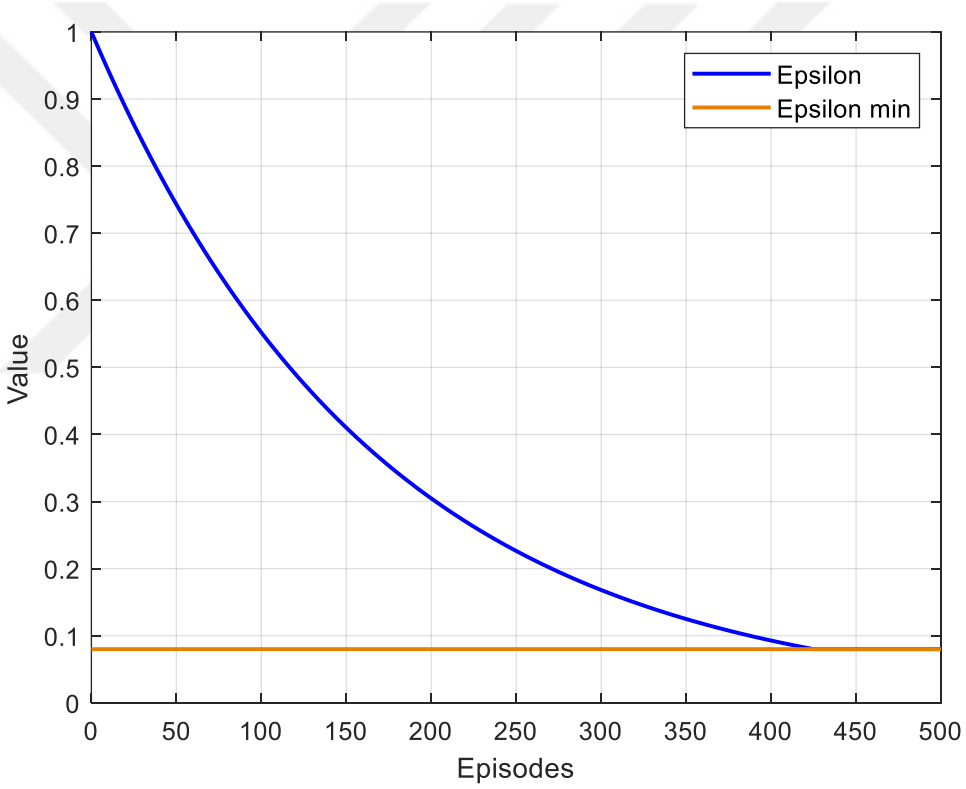

Figure 4.5 : Decay of randomness in exploration exploitation trade-off ($\boldsymbol{\varepsilon_t}$).

4.2.1 Learning environment parameters

Table 4.2 shows the training setup used for the training of DRL based OER control of the PEMFC system. The chosen parameters represent a compromise among learning stability, convergence velocity, and control performance in transient operational scenarios.

The training takes place over 500 episodes, each lasting 8 seconds. This episode length is chosen because it is long enough to show both fast transient dynamics and short-term steady-state behavior of the air supply subsystem, which is important for OER regulation. A sampling time (t_s) of 0.01 s is used to match the learning environment with the compressor and air-path dynamics bandwidth. This lets the agent see high-resolution system responses without putting too much strain on the computer.

A discount factor (γ) of 0.99 is chosen to focus on long-term performance while still taking into account actions that need to be taken right away. This choice makes the agent weigh short-term tracking accuracy against long-term system stability and efficiency. The learning rate (α) is set to 2×10^{-4} which makes the parameter update more stable and less likely to cause oscillatory learning behavior. This is especially important in safety-critical energy systems like PEMFCs.

A soft update factor (τ) of 10^{-3} controls target network updates. This makes sure that the online and target networks sync up slowly. This slow update system helps learning stay stable by stopping sudden changes in the target values during training. The mini-batch size is 256, which keeps the gradient variance low while still allowing for enough randomness to explore the state-action space effectively. Also, a warm-up fraction of 10% is used, during which experience is stored in the replay buffer before policy updates start. This prevents biased learning from not having enough data.

The neural network is fully connected, with hidden layers that have 64, 64, and 32 neurons each. This configuration is chosen to give the PEMFC air system's nonlinear dynamics enough representational power without adding too many parameters, which could cause overfitting or slow convergence.

Table 4.2 : Training parameters.

Parameter	Value
Training episodes	500
Episode duration	8 sec
Sampling time (t_s)	0.01 sec
Discount factor (γ)	0.99
Learning rate (α)	2×10^{-4}
Soft update factor (τ)	10^{-3}
Mini-batch size	256
Warm-up fraction	10%
Network architecture	64-64-32

4.2.2 Reward design

The reward function is made up of several weighted terms that take into account both control goals and physical limits and summed up to obtain total reward.

The main tracking reward imposes penalties on deviations between the measured and reference OER, which shown in equation 4.4, through a quadratic penalty function as in equation 4.5.

$$\lambda_{err} = \lambda_{ref} - \lambda_{measured} \tag{4.4}$$

$$r_{track} = -W_{track}\left(\frac{|\lambda_{err}|}{\lambda_{err,max}}\right)^2 \tag{4.5}$$

The quadratic formulation enforces increasingly severe penalties as the amount of the error increases.

In transient phases, the agent receives rewards for diminishing the magnitude of the tracking error with equation 4.6.

$$r_{recovery} = W_{recovery}\frac{|\lambda_{err}(t-1) - \lambda_{err}(t)|}{\Delta\lambda_{err,max}} \tag{4.6}$$

Asymmetric penalties are imposed for overshoot and undershoot conditions beyond the settling region, proportional to the tracking error. Undershoot reward is assigned to r_{us} a greater weight with the W_{us} weighting coefficient than overshoot as r_{os}, represented by the W_{os} weighting coefficient, to emphasize the greater severity of oxygen starvation relative to excess air supply in PEMFC systems.

Safety limitations on the OER are implemented via a penalty framework featuring soft limits, to r_{limit}, defined as the optimal working range at $\lambda \in [2.0, 2.5]$, and hard limits at $\lambda \in [1.5, 3.0]$, utilizing the W_{limit} limit weighting coefficient. Breaking of strict limitations result in heightened penalties to effectively discourage operation under potentially damaging conditions [28].

To prevent steady-state offset accumulation, an integral error penalty is incorporated described as in equation 4.7.

$$r_{integral} = -W_{integral}\left(\int |\lambda_{err}|\right)^2 \tag{4.7}$$

A multi-tiered settlement bonus rewards the agent for keeping the tracking error within more stringent tolerance bands. The settlement reward is calculated as in equation 4.8.

$$r_{settle} = W_{settle}.\begin{cases} 1.00, |\lambda_{err}| < 0.5\% \\ 0.75, |\lambda_{err}| < 1.0\% \\ 0.50, |\lambda_{err}| < 2.0\% \\ 0.25, |\lambda_{err}| < 5.0\% \\ 0, otherwise \end{cases} \tag{4.8}$$

An oscillation penalty is implemented to avoid chattering behavior during steady-state operation with the help of equation 4.9 when error sign reversal occurs within the settling region.

$$r_{osc} = -W_{osc} \tag{4.9}$$

A penalty on action changes is included to assist with smooth control signals and minimize actuator fatigue with equation 4.10.

$$r_{act} = -W_{act}\left(\frac{\left(\frac{\Delta u}{u_{range}}\right)}{0.5}\right)^2 \left(0.3 + 0.7e^{-5\lambda_{err}}\right) \tag{4.10}$$

This adaptive weighting, based on tracking error, eases the smoothness constraint when serious tracking errors require strong control actions.

Eventually, the final reward is calculated as sum of individual reward components as in equation 4.11.

$$r_{total} = r_{track} + r_{recovery} + r_{us} + r_{os} + r_{limit} + r_{integral} + r_{settle} + r_{osc} + r_{act} \tag{4.11}$$

The reward weights are shown in Table 4.3. The tracking error weight W_{track} emphasizes the need for accurate control of the OER. As mentioned previously, undershoot coefficient is selected bigger than overshoot coefficient since undershoot is directly related with aging of the system. Although the action coefficient attains the highest value, it should be noted that it is active only when the measured OER lies outside the settling band; therefore, its contribution during the agent's operation is limited.

Table 4.3 : Reward weights.

Parameter	Value
W_{track}	50
$W_{recovery}$	40
W_{us}	30
W_{os}	15
W_{limit}	40
$W_{integral}$	30
W_{settle}	50
W_{osc}	15
W_{act}	75

In general, the chosen training parameters and reward weights are meant to make sure that learning is strong while also clearly encoding performance, safety, and efficiency requirements that are important for PEMFC OER control.

The weighting coefficients in the reward function were not established via strict analytical equalization of long-term expected returns; rather, a systematic empirical methodology was employed. Initially, all reward components were normalized to a common scale within the interval $[-1,1]$ to ensure numerical consistency and to prevent any single term from dominating the learning process due to magnitude differences.

After this normalization, the final reward shaping was done through trial-and-error iterative tuning, based on how the agent acted during training. In this process, particular attention was given to how each reward component influences the agent over its entire interaction horizon, rather than focusing solely on instantaneous effects.

It is important to emphasize that each reward term is activated under specific operating conditions and contributes differently depending on the system state. Therefore, the weighting strategy implicitly accounts for the temporal distribution and contextual relevance of these rewards. Some parts have a bigger impact on transient behavior, while others have a bigger impact on steady-state performance or safety-critical situations. So, the chosen weights show a balance not only in size but also in how they affect the agent's life as a whole, making sure that the policy learned meets both performance and safety goals.

4.3 Testing

A step change test scenario is established in which the load current varies between 105 and 215 A [11]. Consequently, the OER reference varies between 2.15 and 2.475. Two agents are trained with direct DQN and PI-augmented DQN architectures for previously defined learning parameters as in Table 4.4.

It is crucial to note that the final results reported in this study were not derived from a single, fixed reward formulation in order to give an open account of the controller development process. Rather, they are the result of a reward shaping process that involved the evaluation and iterative improvement of multiple intermediate designs.

In the initial stage, a relatively simple reward structure was employed, primarily focusing on instantaneous OER tracking performance. But, as shown in Figure 4.6, this formulation led to a clear steady-state tracking error, which was shown by a consistently high OER error term. The agent could adapt to changes, but it didn't have any reason to get rid of long-term differences from the reference.

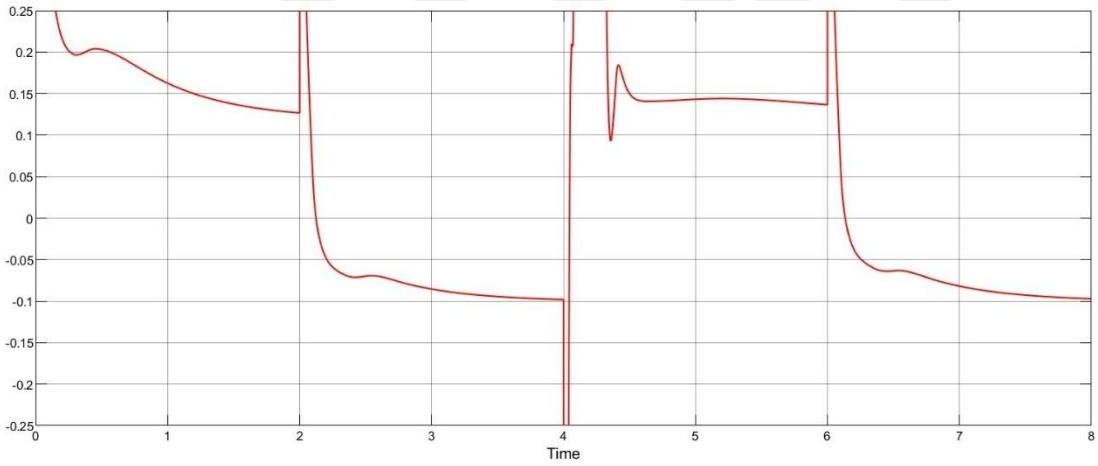

Figure 4.6 : High steady state OER error without integral term.

To fix this problem, the reward function was improved by adding an integral error term and putting more focus on tracking references. This change made the agent take into account deviations that had built up over time, which greatly lowered the steady-state error and improved the overall tracking accuracy. The improvement is clear in the response, where the OER gets closer to its reference value.

A second limitation observed in the early reward designs was related to control effort behavior. As shown in Figure 4.7, the compressor current exhibited pronounced

oscillations when no explicit penalty on the control action was included in the reward function. This led to physically undesirable actuator behavior, despite acceptable tracking performance.

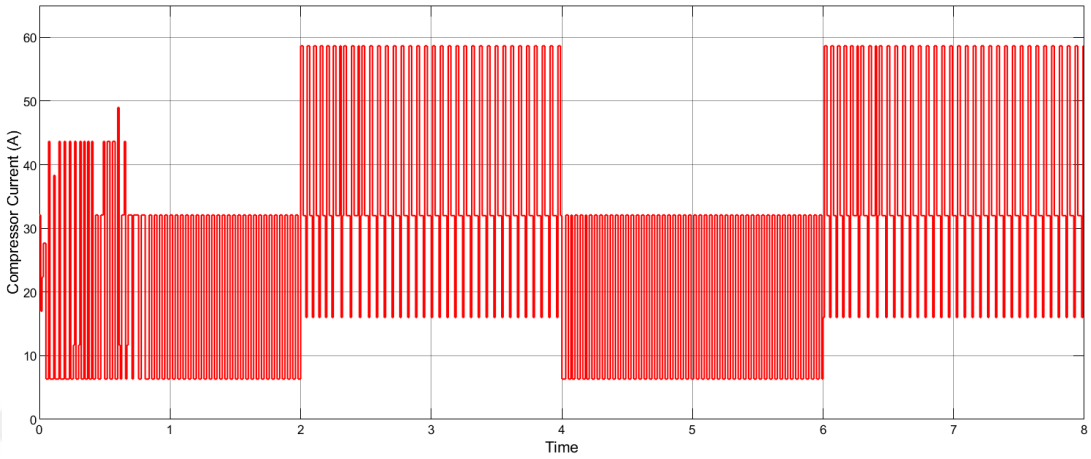

Figure 4.7 : Oscillatory compressor current.

To fix this problem, an extra reward term was added that punished too much change in the control action. The agent learned smoother control policies by making it clear that rapid changes were not wanted. This made the compressor current profile more stable without reducing tracking performance.

These intermediate results demonstrate that the final control performance is a direct consequence of progressive reward refinement. Presenting these less successful initial configurations not only highlights the challenges inherent in reward design for reinforcement learning but also substantiates the effectiveness of the proposed approach in achieving both accurate tracking and physically consistent control behavior.

The final reward evolution over training episodes for the Direct DQN and the DQN augmented with a PI controller (PI–DQN) with configuration 1 is illustrated in Figure 4.8.

At the initial stage of training, both agents exhibit strongly negative reward values, which is expected due to the predominantly exploratory behavior and the lack of an informed policy. However, the Direct DQN starts from a significantly lower reward level and shows a slower recovery phase. This behavior indicates that the agent must simultaneously learn the control structure and compensate for the system dynamics without any prior guidance, resulting in a prolonged transient learning period.

Conversely, the embedded PI controller provides a stabilizing baseline behavior, allowing the reinforcement learning agent to focus primarily on corrective and fine-tuning actions rather than learning the entire control task from scratch which results higher initial reward during warmup phase.

As training progresses, both approaches converge toward higher reward levels; however, the PI–DQN consistently achieves a higher steady-state reward taking into account that the meaining of the action in rewards section changes for these two different structures with extra two observation states.

Overall, the comparison highlights that while the Direct DQN is capable of learning a viable control policy, as well as hybrid architecture.

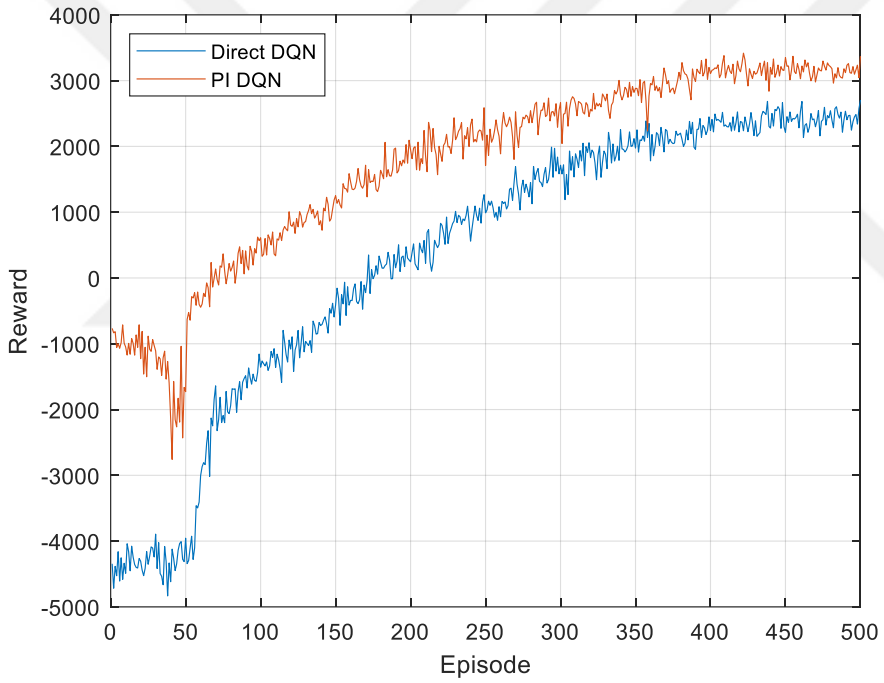

Figure 4.8 : Training rewards.

Although the exploration rate ε_t gradually decreases and reaches its minimum value, the episode-level Q-value does not necessarily converge to a constant value, as in Figure 4.9. Instead, the observed monotonic increase reflects the improvement of the expected cumulative return as the policy becomes more deterministic and consistently selects higher-reward actions. At the beginning, both of the Q Values behold near zero caused by warmup episodes. The value function's bootstrapped nature and the use of function approximation are what mostly cause the episode Q-value to go up. These

two things let incremental policy improvements move through the Q estimates from one episode to the next.

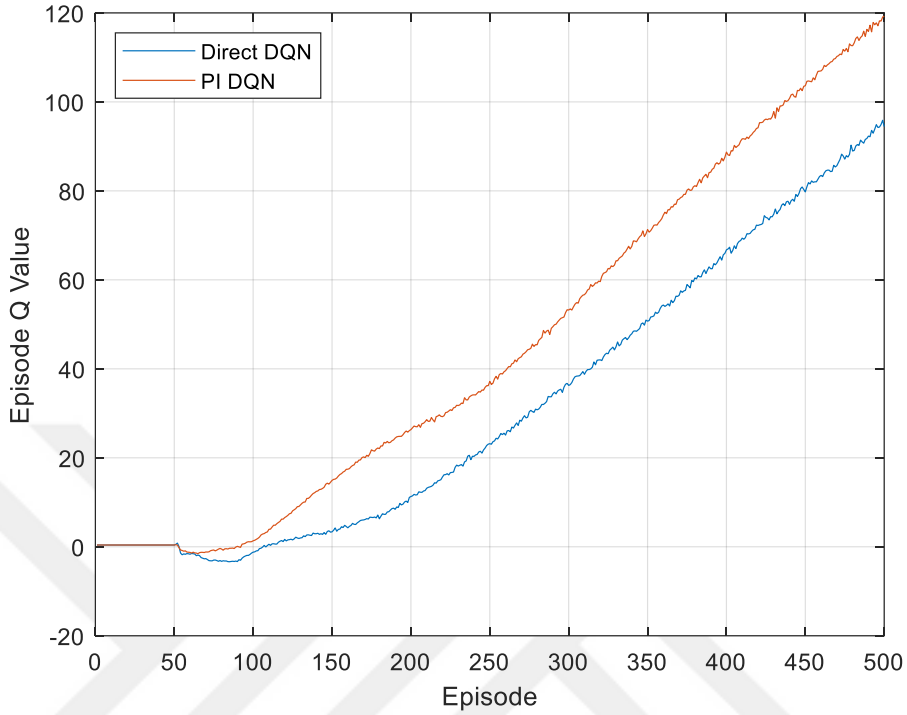

Figure 4.9 : Training Q values.

The corresponding control signal in Figure 4.10 indicates that the integration of DQN with the PI controller results in significant fluctuations in compressor current. The Direct DQN controller exhibits slight fluctuations in the compressor current due to its utilization of a discrete action space, resulting in step-like corrections instead of smooth changes.

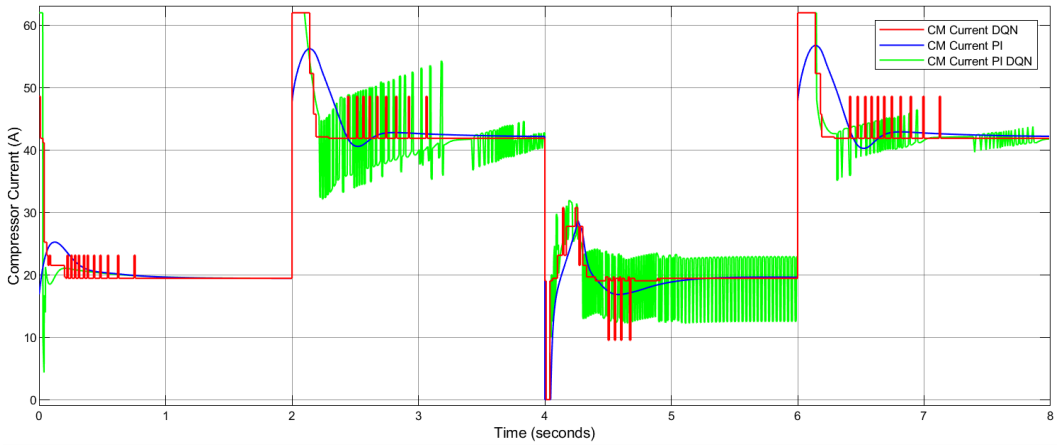

Figure 4.10 : Compressor current.

The measured OER are illustrated in Figure 4.11. DQN-based controllers have a more rapid response. Furthermore, it exhibits superior overshoot and undershoot performance due to its nonlinear control capabilities. Conversely, the PI controller exhibits a reduced steady-state error due to the integral component.

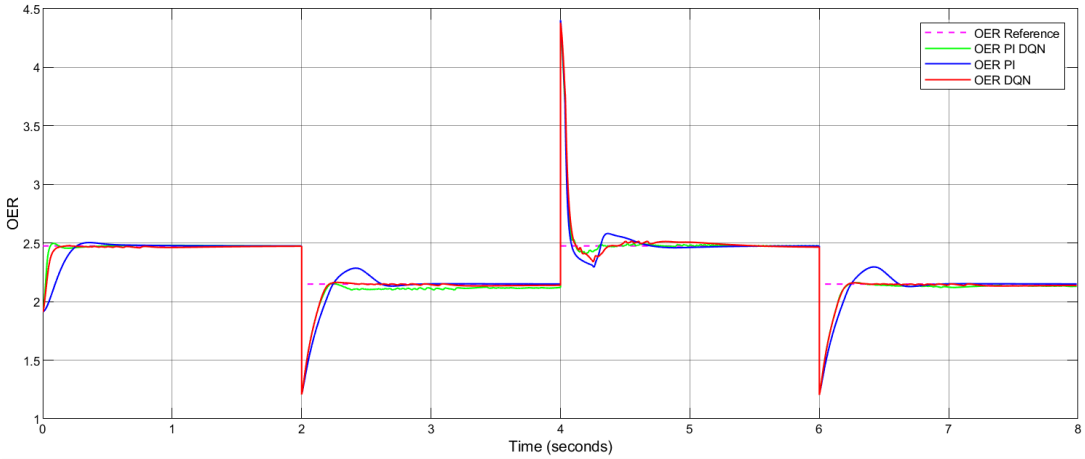

Figure 4.11 : Measured OER.

Figure 4.12 illustrates that DQN-based controllers have higher performance in stabilizing OER in shorter duration, although the PI controller outperforms at steady-state responsiveness with small amount of steady-state error.

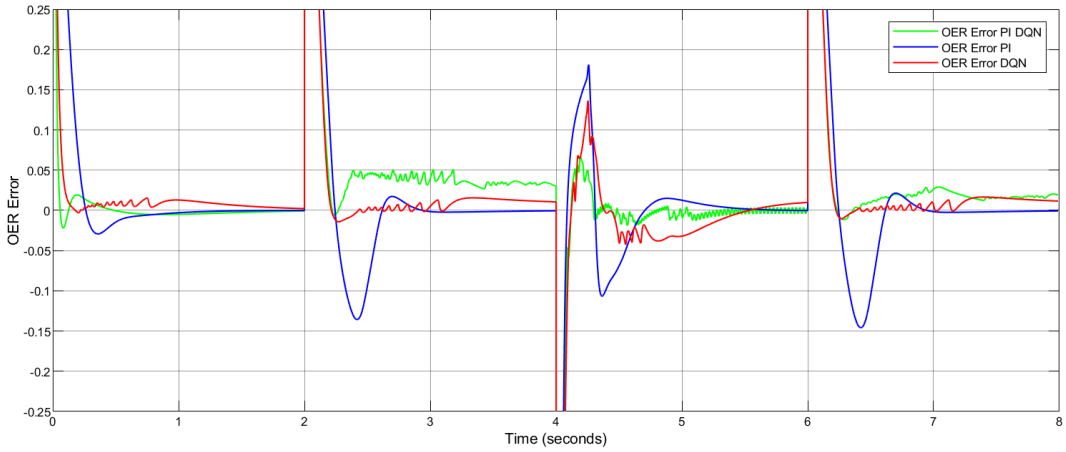

Figure 4.12 : OER error.

Figure 4.13 illustrates that the PI controller stabilizes net power production effectively. Nonetheless, oscillations occur with DQN-based controllers, particularly when DQN is integrated with a PI controller. These variations are directly associated with the compressor current.

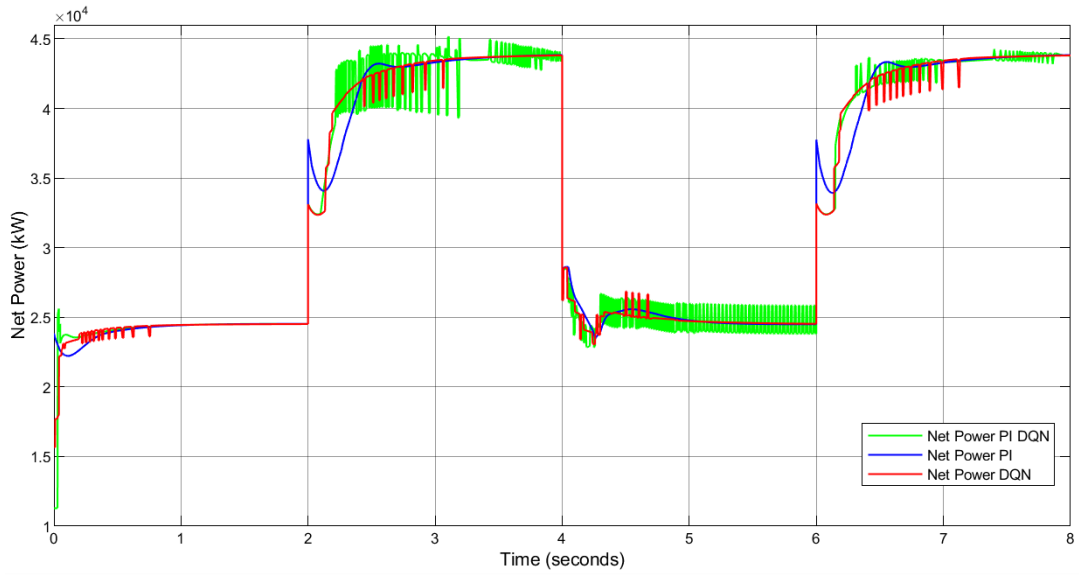

Figure 4.13 : Net power output.

It is possible to make the system response more aggressive or more robust by calibrating PI controller coefficients accordingly but in that case system response has big amount of overshoot/undershoot or has longer settling time, as can be seen in Figure 4.14.

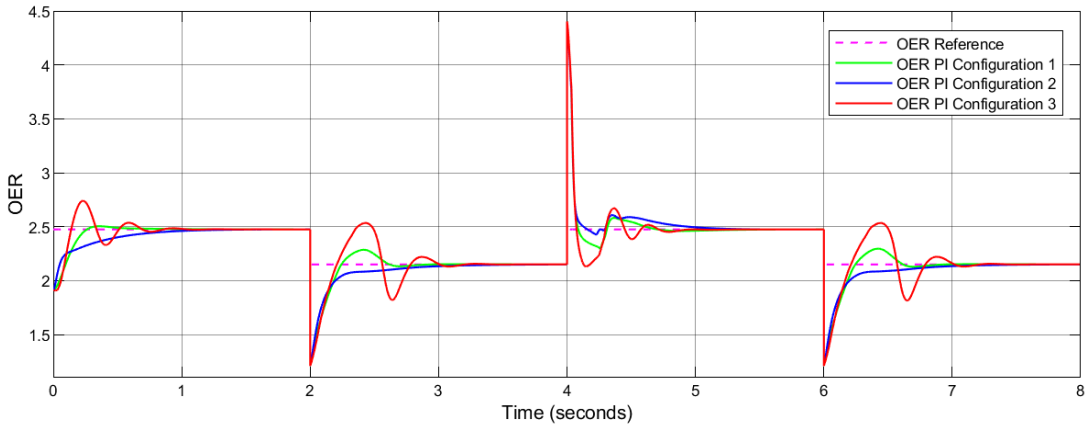

Figure 4.14 : Measured OER with classical controllers.

On the other hand, DQN based controllers has both small amount of overshoot/undershoot and shorter settling time at the same time. Nonetheless, DQN based controllers fall behind of PI controller in the scope of steady-state response performance with higher amount of steady-state error caused by lack of strong enough integral term.

Another training has been conducted with defective PI controller configurations 4 and 5, with identical training parameters.

Figure 4.15 presents the evolution of the cumulative reward during training for the DQN agent augmented with two mistuned PI controllers, corresponding to an aggressive and a robust tuning configuration. In both cases, the training process starts with negative reward values, reflecting the combined effect of exploration and the performance degradation introduced by the intentionally unbalanced PI parameters.

During the early training phase, the configuration associated with the mistuned aggressive PI exhibits a sharper and deeper reward drop compared to the mistuned robust case. This behavior indicates that the aggressive tuning amplifies control actions under exploration, leading to pronounced transient deviations and consequently larger penalties in the reward signal.

As training progresses, both configurations show a consistent and monotonic improvement in reward, eventually converging toward a similar asymptotic performance level. This convergence demonstrates that the reinforcement learning agent is capable of compensating for significant deficiencies in the underlying PI controller by learning corrective actions that dominate the closed-loop behavior. Notably, the long-term reward trajectories of the two configurations become nearly indistinguishable, indicating that the learned policy effectively reduces sensitivity to the initial PI tuning characteristics.

The results highlight an important robustness property of the proposed PI–DQN framework. This observation confirms that the hybrid control structure can tolerate substantial controller mismatch and still achieve stable and high-reward operation after sufficient training.

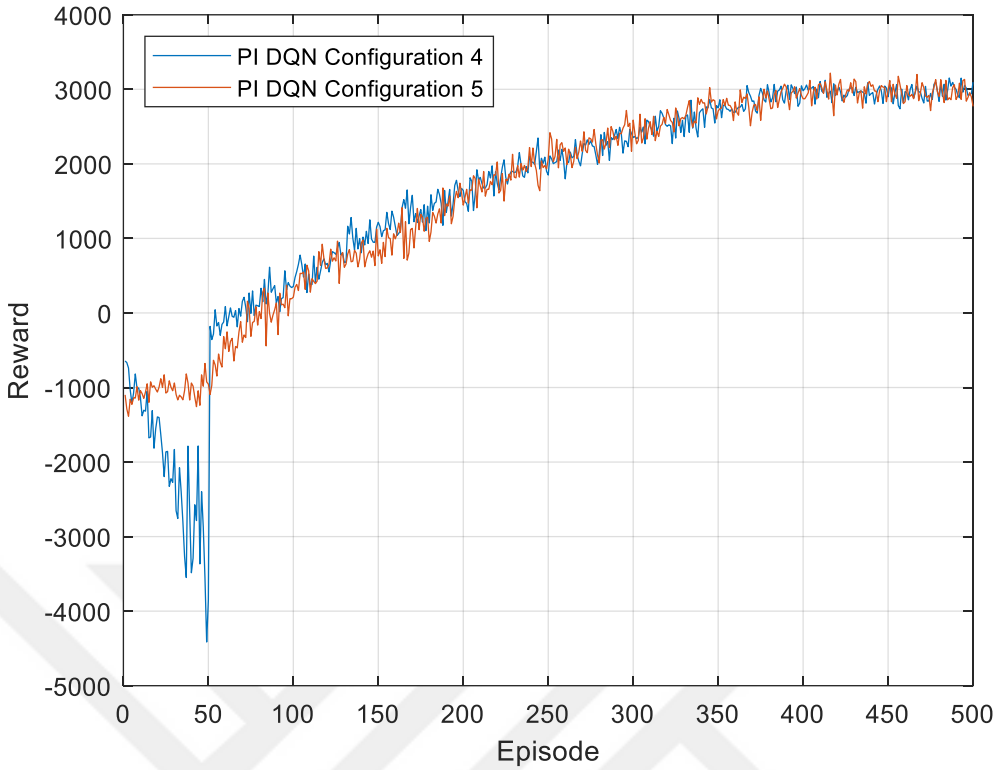

Figure 4.15 : Training reward with mistuned PI controllers.

Figure 4.16 illustrates the evolution of the episode-level Q-values during training for the PI–DQN framework under mistuned aggressive and mistuned robust PI controller configurations. In both cases, the Q-values exhibit a monotonic increase over the training horizon, indicating continuous improvement in the estimated expected cumulative return as the learning process progresses.

At the early stages of training, the Q-values remain close to zero for both configurations because of warmup episodes.

As training advances, the aggressive configuration shows a slightly faster growth rate in the Q-values compared to the robust configuration. This behavior suggests that the higher control authority associated with the aggressive PI tuning enables the reinforcement learning agent to experience larger state-action value gradients, which accelerates the propagation of reward information through the value function.

Despite the initial difference in growth rate, the two trajectories remain closely aligned throughout most of the training process. The robust configuration follows a similar trend with a marginal delay, indicating that conservative control gains slightly slow down the accumulation of expected future rewards but do not prevent effective value

learning. Toward the later training episodes, both configurations converge to comparable Q-value ranges, implying that the learned policy progressively compensates for the deficiencies introduced by the mistuned PI parameters. This result proves that the PI–DQN architecture maintains stable and consistent value function learning even when the underlying PI controller is significantly unbalanced, highlighting that the reinforcement learning component effectively mitigates controller mismatch and reduces sensitivity to the initial PI tuning in the long term.

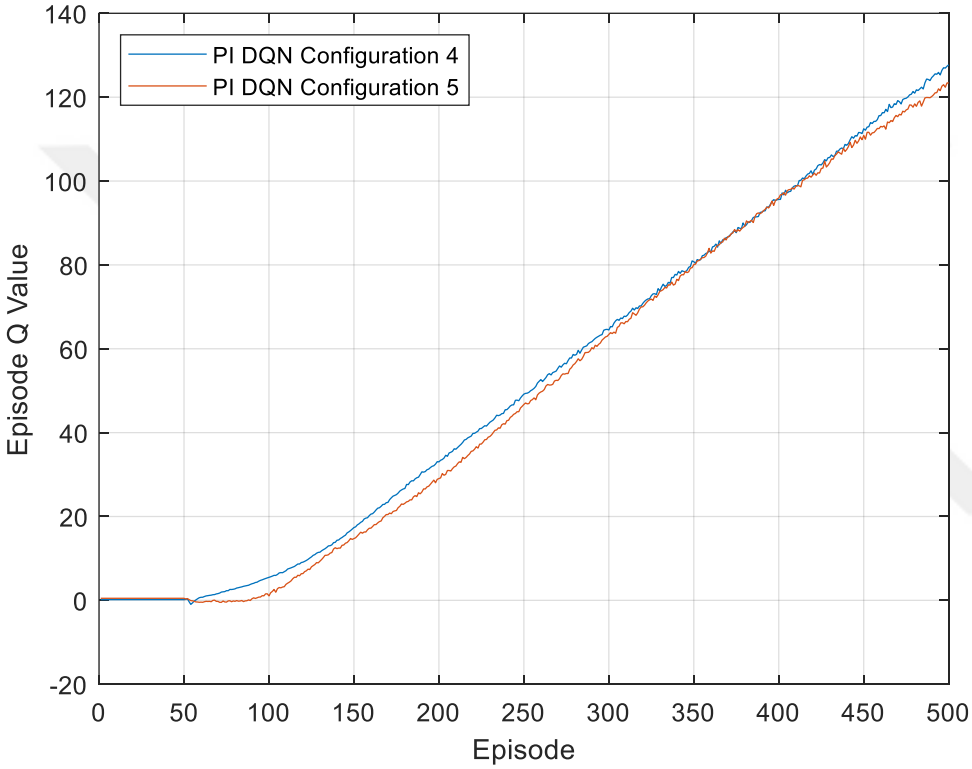

Figure 4.16 1: Training Q values with mistuned PI controllers.

Figure 4.17 clearly illustrates that improper tuning of PI controller gains can result in oscillatory or slower responses. The implementation of DQN enhances system responsiveness and mitigates significant oscillations. Regrettably, a little steady-state fault persists. To mitigate this steady-state errors, the integral gain $W_{integral}$ can be incremented; however, this adjustment results in a slower system response.

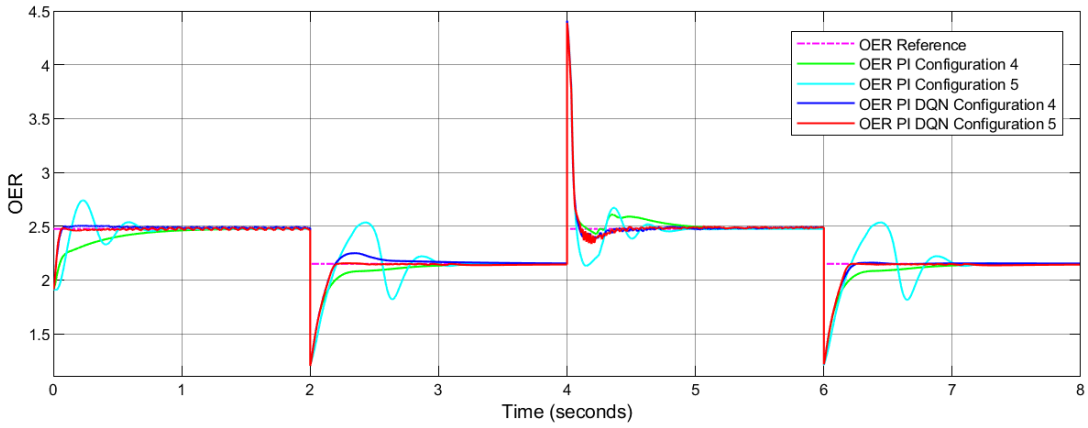

Figure 4.17 : Measured OER with mistuned PI controllers.

The test results reported in Table 4.4 provide a comparative evaluation of Direct DQN and hybrid PI–DQN configurations in terms of OER regulation accuracy and net power performance, where IAE is integral absolute error and SS error is steady-state error. The Direct DQN configuration achieves a relatively low OER IAE of 0.3490 with a moderate steady-state error of 0.011, demonstrating that a purely data-driven policy can maintain acceptable tracking performance without relying on classical control structure. However, its steady-state accuracy remains inferior to several PI-based configurations, indicating limited precision once the system dynamics settle, while considering the average net power output remains within a narrow band across all configurations. This implies that maximizing net power alone does not necessarily coincide with optimal OER regulation, and that aggressive power-oriented control may sacrifice oxygen management quality. On the other hand, aggressive control reduces starvation and prolong system life time.

Introducing the DQN agent as a corrective layer on top of PI controllers generally improves transient performance, as reflected by reduced OER IAE values in most PI–DQN cases compared to their standalone PI counterparts. Notably, PI-DQN configuration 5 achieves the lowest OER IAE with 0.3126, suggesting that the reinforcement learning agent is particularly effective in compensating for suboptimal PI tuning during transient operating conditions. This confirms the hypothesis that DQN is able to correct the behivour when controller parameters are calibrated nonaccurate or system dynamics change over time.

Nevertheless, this improvement does not come without cost. The same configuration exhibits a significantly higher steady-state error with 0.11, which reveals an important

limitation: while the RL agent enhances transient regulation, it may interfere with steady-state convergence when its corrective action is not sufficiently attenuated near equilibrium.

Table 4.4 : Performance comparison.

Configuration	OER IAE (-)	OER SS error (-)	Average net power output (kW)
Direct DQN	0.3490	0.011	33285.03
PI configuration 1	0.47	0.001	33283.96
PI DQN configuration 1	0.3624	0.0108	33306.80
PI configuration 2	0.4640	0.001	33358.88
PI configuration 3	1.1887	0.001	33060.24
PI configuration 4	0.5129	0.001	33280.89
PI DQN configuration 4	0.3543	0.0085	33215.56
PI configuration 5	0.7506	0.001	33324.24
PI DQN configuration 5	0.3126	0.11	33287.64

Table 4.5 presents a timing-oriented performance comparison of the evaluated control configurations, focusing on dynamic response speed and the duration of operation outside predefined OER safety and optimality bounds. In this context, OER violation time represents the cumulative duration during which the oxygen excess ratio lies outside the safe interval of 1.5–3, while OER deviation time corresponds to operation outside the optimal efficiency range of 2–2.5.

The settling time results indicate a clear advantage for reinforcement learning–assisted configurations. While the Direct DQN controller achieves a relatively fast settling time of 0.28 s, several PI–DQN configurations outperform both Direct DQN and classical PI controllers. In particular, PI-DQN configuration 1 exhibits the shortest settling time of 0.19 s, demonstrating that the RL agent effectively accelerates transient convergence when embedded as a corrective layer. In contrast, standalone PI controllers show significantly slower responses, with settling times reaching up to 0.91 s in PI configuration 3, highlighting their limited adaptability under transient conditions.

From a safety standpoint, OER violation time reveals a more nuanced picture. Classical PI controllers show substantial variability, with PI configuration 3 spending as much as 0.35 s outside the safe OER bounds, which can be considered critical from a fuel cell durability perspective. Reinforcement learning–assisted controllers

consistently reduce this metric, with several PI–DQN configurations achieving the minimum observed violation time of 0.14 s. This reduction suggests that the RL agent successfully suppresses extreme OER excursions, thereby contributing to safer operation during load transients.

The most pronounced benefit of the PI–DQN architecture emerges in the OER deviation time metric. Such prolonged operation in suboptimal regions implies efficiency loss even when safety limits are not violated. In contrast, PI–DQN configurations significantly shorten deviation time, with PI DQN configurations 1 and 5 reducing it to approximately 0.44 s and 0.43 s, respectively. This indicates that the RL agent not only mitigates hazardous excursions but also actively drives the system back toward energetically favorable operating conditions.

It is worth noting that the Direct DQN controller, while competitive in settling time, does not achieve the same level of optimal-region adherence as the best PI–DQN cases. This observation reinforces the importance of incorporating structured control knowledge, such as PI regulation, to guide learning-based policies toward consistent and interpretable behavior.

Table 4.5 : Timing performance comparison.

Configuration	Settling time (sec)	OER violation time (sec)	OER deviation time (sec)
Direct DQN	0.28	0.16	0.85
PI configuration 1	0.55	0.17	0.92
PI DQN configuration 1	0.19	0.15	0.44
PI configuration 2	0.44	0.15	0.58
PI configuration 3	0.91	0.35	2.4
PI configuration 4	0.74	0.15	1.2
PI DQN configuration 4	0.24	0.14	0.67
PI configuration 5	0.71	0.18	1.51
PI DQN configuration 5	0.24	0.14	0.43

Overall, the results indicate that PI–DQN cooperation significantly enhances transient OER control by accelerating response speed and reducing both safety-critical OER violations and efficiency-degrading deviations, particularly in cases where the baseline PI controller is imperfectly tuned. At the same time, the degradation in steady-state accuracy observed in certain configurations highlights the need for careful reward shaping and action constraint design to avoid persistent corrective behavior near equilibrium. These findings demonstrate that hybrid control performance is inherently

configuration-dependent and confirm that reinforcement learning is most effective when employed as a complementary corrective mechanism rather than a standalone replacement for classical PI control in PEMFC air management systems, where durability and efficiency must be simultaneously preserved.

5. CONCLUSIONS

This study investigated the integration of DRL, specifically the DQN algorithm, with a conventional PI controller for the control of the OER in a PEMFC air supply system. The primary objective was to evaluate whether reinforcement learning can enhance transient performance and compensate for structural limitations of classical control strategies under varying load conditions.

The simulation results demonstrate that the standalone PI controller provides the smoothest control response in terms of output continuity and steady-state stability. However, this smoothness is achieved at the expense of transient performance, particularly under abrupt load variations where slower dynamic adaptation is observed. In contrast, the DQN-based controller exhibits rapid responsiveness to changing operating conditions, reflecting its capacity to learn adaptive control policies directly from interaction with the environment. Nevertheless, this improved adaptability is accompanied by oscillatory behavior, primarily attributable to the discretized action space inherent in the DQN formulation. Since compressor motor current adjustments are restricted to predefined discrete increments, the agent cannot produce infinitesimally small corrections near equilibrium, which leads to persistent fluctuations around the reference value.

Moreover, the multi-objective reward structure introduces an additional layer of complexity. The simultaneous optimization of tracking accuracy, efficiency, and safety-related OER constraints creates competing incentives within the learning process. Near reward boundaries, the agent may alternate between policies that individually favor different objectives, thereby producing local oscillations. This observation highlights the sensitivity of reinforcement learning performance to reward shaping and underscores the necessity of carefully balancing competing control objectives in safety-critical energy systems.

A key contribution of this study lies in demonstrating that reinforcement learning, when embedded within a hybrid PI–DQN architecture, offers functionality beyond simple error compensation. The reinforcement learning component effectively acts as

an adaptive supervisory layer that enhances or corrects the behavior of the baseline PI controller. Notably, the PI–DQN structure successfully restored acceptable control performance even when the PI gains were intentionally detuned or degraded. This finding suggests that RL can serve as a resilience mechanism, mitigating the effects of parameter drift, modeling inaccuracies, or aging-related performance degradation. In practical PEMFC applications, where plant dynamics vary due to temperature changes, membrane hydration levels, compressor wear, and long-term aging, such adaptive capability is particularly valuable.

The results further indicate that reinforcement learning does not necessarily replace classical control; rather, it augments it. The PI controller provides structural stability and interpretability, while the DQN contributes adaptive fine-tuning and compensation. This synergy reduces the exploration burden of the learning agent and accelerates convergence by embedding prior control knowledge into the system. Therefore, hybridization emerges as a pragmatic pathway toward deploying learning-based control in real-time fuel cell systems without sacrificing robustness.

Despite these promising findings, several limitations remain. The discrete action formulation inherently restricts steady-state precision. Continuous-action RL methods, such as actor–critic frameworks, may enable smoother compressor current modulation and improved steady-state accuracy. Additionally, the fixed weighting of reward components constrains adaptability. Future research may investigate adaptive or performance-dependent reward weighting mechanisms, allowing the controller to dynamically prioritize efficiency, durability, or tracking performance depending on operating conditions.

Further investigation is also required to assess robustness against parameter uncertainties, sensor noise, membrane aging effects, and unmodeled nonlinearities. While simulation results are encouraging, real-time implementation introduces computational and hardware constraints that must be carefully evaluated. Optimization of network size, inference latency, and embedded hardware compatibility will be essential for practical deployment.

Finally, future studies may explore broader hybrid control configurations, integrating reinforcement learning with alternative classical strategies such as MPC, LQR, or fuzzy logic control. Comparative analyses across these architectures would provide

deeper insight into the trade-offs between adaptability, interpretability, computational complexity, and control performance in PEMFC air management systems.

In conclusion, this work demonstrates that reinforcement learning, particularly within a hybrid PI–DQN framework, represents a viable and promising approach for enhancing OER control in PEMFC systems. Rather than serving as a standalone replacement for classical controllers, DRL functions most effectively as an adaptive augmentation mechanism, capable of improving transient response, compensating for parameter degradation, and increasing overall system resilience in dynamically varying operating environments.

REFERENCES

[1] **Ahn, J.-W., & Choe, S.-Y.** (2008). Coolant controls of a PEM fuel cell system. *Journal of Power Sources*, *179*(1), 252–264. https://doi.org/10.1016/j.jpowsour.2007.12.066

[2] **Aydin, M. I., & Dincer, I.** (2022). A life cycle impact analysis of various hydrogen production methods for public transportation sector. *International Journal of Hydrogen Energy*, *47*(93), 39666–39677. https://doi.org/10.1016/j.ijhydene.2022.09.125

[3] **Bellman, R.** (1957). A Markovian Decision Process. *Indiana University Mathematics Journal*, *6*(4), 679–684. https://doi.org/10.1512/iumj.1957.6.56038

[4] **Chen, J., Liu, Z., Wang, F., Ouyang, Q., & Su, H.** (2018). Optimal Oxygen Excess Ratio Control for PEM Fuel Cells. *IEEE Transactions on Control Systems Technology*, *26*(5), 1711–1721. https://doi.org/10.1109/TCST.2017.2723343

[5] **Deng, Z., Chen, M., Chen, Q., & Wang, H.** (2024). Oxygen starvation control of proton exchange membrane fuel cell through fusion control strategy. *Applied Soft Computing*, *159*, 111655. https://doi.org/10.1016/j.asoc.2024.111655

[6] **Gerard, M., Poirot-Crouvezier, J.-P., Hissel, D., & Pera, M.-C.** (2010). Oxygen starvation analysis during air feeding faults in PEMFC. *International Journal of Hydrogen Energy*, *35*(22), 12295–12307. https://doi.org/10.1016/j.ijhydene.2010.08.028

[7] **Hägglund, T.** (2019). The one-third rule for PI controller tuning. *Computers & Chemical Engineering*, *127*, 25–30. https://doi.org/10.1016/j.compchemeng.2019.03.027

[8] *https://en.wikipedia.org/wiki/Fuel_cell*.

[9] **İstanbul Technical University,** *EBT 526E Fuel Cell Lecture Notes*. (2024).

[10] **Jia, C., Zhou, J., He, H., Li, J., Wei, Z., Li, K., & Shi, M.** (2023). A novel energy management strategy for hybrid electric bus with fuel cell health and battery thermal- and health-constrained awareness. *Energy*, *271*, 127105. https://doi.org/https://doi.org/10.1016/j.energy.2023.127105

[11] **Kendir, F., & Kumbasar, T.** (2024). Design of 2DOF control system fused with artificial intelligence for power enhancement and mitigation of degradation in fuel cell systems. *Expert Systems with Applications*, *249*, 123632. https://doi.org/10.1016/j.eswa.2024.123632

[12] **Kordesch, K. V., & Simader, G. R.** (1996). *Fuel Cells*. Wiley. https://doi.org/10.1002/352760653X

[13] **Larminie, J., & Dicks, A.** (2003). *Fuel Cell Systems Explained*. Wiley. https://doi.org/10.1002/9781118878330

[14] **LeCun, Y., Bengio, Y., & Hinton, G.** (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

[15] **Li, J., & Yu, T.** (2021). A new adaptive controller based on distributed deep reinforcement learning for PEMFC air supply system. *Energy Reports*, *7*, 1267–1279. https://doi.org/10.1016/j.egyr.2021.02.043
[16] **Li, J., & Yu, T.** (2021). Sensors Integrated Control of PEMFC Gas Supply System Based on Large-Scale Deep Reinforcement Learning. *Sensors*, *21*(2), 349. https://doi.org/10.3390/s21020349
[17] **Li, J., Yu, T., & Yang, B.** (2021). Coordinated control of gas supply system in PEMFC based on multi-agent deep reinforcement learning. *International Journal of Hydrogen Energy*, *46*(68), 33899–33914. https://doi.org/10.1016/j.ijhydene.2021.07.009
[18] **Li, W., Li, S., Du, C., Xu, Y., Xin, Q., & Yan, F.** (2025). A deep reinforcement learning control method for PEMFC thermal management and air supply system. *Applied Thermal Engineering*, *279*, 128030. https://doi.org/10.1016/j.applthermaleng.2025.128030
[19] **Liu, Y.-J., Gong, M., Liu, L., Tong, S., & Chen, C. L. P.** (2021). Fuzzy Observer Constraint Based on Adaptive Control for Uncertain Nonlinear MIMO Systems With Time-Varying State Constraints. *IEEE Transactions on Cybernetics*, *51*(3), 1380–1389. https://doi.org/10.1109/TCYB.2019.2933700
[20] **Liu, Y.-J., Gong, M., Tong, S., Chen, C. L. P., & Li, D.-J.** (2018). Adaptive Fuzzy Output Feedback Control for a Class of Nonlinear Systems With Full State Constraints. *IEEE Transactions on Fuzzy Systems*, *26*(5), 2607–2617. https://doi.org/10.1109/TFUZZ.2018.2798577
[21] **Lu, Z., & Yan, Y.** (2024). Temperature Control of Fuel Cell Based on PEI-DDPG. *Energies*, *17*(7), 1728. https://doi.org/10.3390/en17071728
[22] **Meng, J., Guo, Q., Zhang, X., Lin, J., & Yue, M.** (2025). A robust adaptive control strategy for oxygen excess ratio regulation in fuel cell systems with reduced control effort. *ISA Transactions*, *167*, 1495–1505. https://doi.org/10.1016/j.isatra.2025.09.032
[23] **Mnih, & Volodymyr**. (2013). Playing atari with deep reinforcement learning. *ArXiv Preprint*.
[24] **Ng, Andrew Y, Daishi Harada, & Stuart Russell**. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, *99*, 278–287.
[25] **Nian, R., Liu, J., & Huang, B.** (2020). A review On reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, *139*, 106886. https://doi.org/10.1016/j.compchemeng.2020.106886
[26] **Niknezhadi, A., Allué-Fantova, M., Kunusch, C., & Ocampo-Martínez, C.** (2011). Design and implementation of LQR/LQG strategies for oxygen stoichiometry control in PEM fuel cells based systems. *Journal of Power Sources*, *196*(9), 4277–4282. https://doi.org/10.1016/j.jpowsour.2010.11.059
[27] **Özdemir, T. H., Şen, C., & Yüksel, L.** (2023). Proton Exchange Membrane Fuel Cell System Mathematical Modeling on MATLAB Simulink Environment. *2023 14th International Conference on Electrical and Electronics Engineering (ELECO)*, 1–5. https://doi.org/10.1109/ELECO60389.2023.10415931

[28] **Pukrushpan, J. T., Peng, H., & Stefanopoulou, A. G.** (2004). Control-Oriented Modeling and Analysis for Automotive Fuel Cell Systems. *Journal of Dynamic Systems, Measurement, and Control*, *126*(1), 14–25. https://doi.org/10.1115/1.1648308

[29] **Puterman, M. L.** (1994). *Markov Decision Processes*. Wiley. https://doi.org/10.1002/9780470316887

[30] **Richard S. Sutton, & Andrew G. Barto**. (2018). *Reinforcement Learning : An Introduction* ((2nd ed.)).

[31] **Sankar, K., & Jana, A. K.** (2018). Dynamics and Estimator-Based Nonlinear Control of a PEM Fuel Cell. *IEEE Transactions on Control Systems Technology*, *26*(3), 1124–1131. https://doi.org/10.1109/TCST.2017.2695165

[32] **Springer, T. E., Zawodzinski, T. A., & Gottesfeld, S.** (1991). Polymer Electrolyte Fuel Cell Model. *Journal of The Electrochemical Society*, *138*(8), 2334–2342. https://doi.org/10.1149/1.2085971

[33] **Sürer, M. G., & Arat, H. T.** (2022). Advancements and current technologies on hydrogen fuel cell applications for marine vehicles. *International Journal of Hydrogen Energy*, *47*(45), 19865–19875. https://doi.org/10.1016/j.ijhydene.2021.12.251

[34] **Talj, R., Ortega, R., & Astolfi, A.** (2011). Passivity and robust PI control of the air supply system of a PEM fuel cell model. *Automatica*, *47*(12), 2554–2561. https://doi.org/10.1016/j.automatica.2011.08.028

[35] **Vahidi, A., Stefanopoulou, A., & Peng, H.** (2006). Current Management in a Hybrid Fuel Cell Power System: A Model-Predictive Control Approach. *IEEE Transactions on Control Systems Technology*, *14*(6), 1047–1057. https://doi.org/10.1109/TCST.2006.880199

[36] **Watkins, C. J. C. H., & Dayan, P.** (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

[37] **Wolf Vielstich, Arnold Lamm, & Hubert A. Gasteiger**. (2003). *Handbook of Fuel Cells: Fundamentals, Technology and Applications* (Vol. 3).

[38] **Yeetsorn, R., Maiket, Y., & Kaewmanee, W.** (2020). The observation of supercapacitor effects on PEMFC–supercapacitor hybridization performance through voltage degradation and electrochemical processes. *RSC Advances*, *10*(22), 13100–13111. https://doi.org/10.1039/D0RA00468E

[39] **Yildirim, B., Gheisarnejad, M., Özdemir, M. T., & Khooban, M. H.** (2024). Multi-agent fuzzy Q-learning-based PEM fuel cell air-feed system control. *International Journal of Hydrogen Energy*, *75*, 354–362. https://doi.org/10.1016/j.ijhydene.2024.02.129

[40] **Zhang, W., Shen, X., Zhang, H., Yin, Z., Sun, J., Zhang, X., & Zou, L.** (2024). Feature importance measure of a multilayer perceptron based on the presingle-connection layer. *Knowledge and Information Systems*, *66*(1), 511–533. https://doi.org/10.1007/s10115-023-01959-7

[41] **Zhou, J., Liu, J., Su, Q., Feng, C., Wang, X., Hu, D., Yi, F., Jia, C., Fan, Z., & Jiang, S.** (2022). Heat Dissipation Enhancement Structure Design of Two-Stage Electric Air Compressor for Fuel Cell Vehicles Considering Efficiency Improvement. *Sustainability*, *14*(12), 7259. https://doi.org/10.3390/su14127259

CURRICULUM VITAE

Name Surname **:** Emir Oğuz

EDUCATION **:**

- **B.Sc.** **:** 2022**,** Yıldız Technical University, Faculty of Electrical and Electronics Engineering, Control and Automation Engineering.

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2026 – Present: Senior Systems Engineer, AVL Türkiye Research and Engineering
- 2022 – 2026: Systems Engineer, AVL Türkiye Research and Engineering
- 2021: Intern, University of Applied Sciences Emden/Leer
- 2021: Intern, TÜBİTAK Marmara Research Center

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Oğuz, E.,** & Onat, A. (2026). Fuel Cell Air Supply System Control Using Deep Q-Network. In 2026 13th International Conference on Electrical and Electronics Engineering (ICEEE). IEEE.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Oğuz, E.,** & Şen, C. (2026). System-Level Modelling of a Hydrogen Internal Combustion Engine with Balance of Plant Components: Performance and Emissions Analysis. In 2026 13th International Conference on Electrical and Electronics Engineering (ICEEE). IEEE.
- Şen, C., **Oğuz, E.,** & Yüksel, L**.** (2025, September). Modelling of a Solid Oxide Electrolyzer Cell and Its Balance of Plant System. In 2025 12th International Conference on Electrical and Electronics Engineering (ICEEE) (pp. 300-304). IEEE.
- Sariçay, B., **Oğuz, E.,** & Kiliç, M. (2024, April). Dynamic Modelling of a 2-kW Solid Oxide Fuel Cell System. In 2024 11th International Conference on Electrical and Electronics Engineering (ICEEE) (pp. 249-253). IEEE.