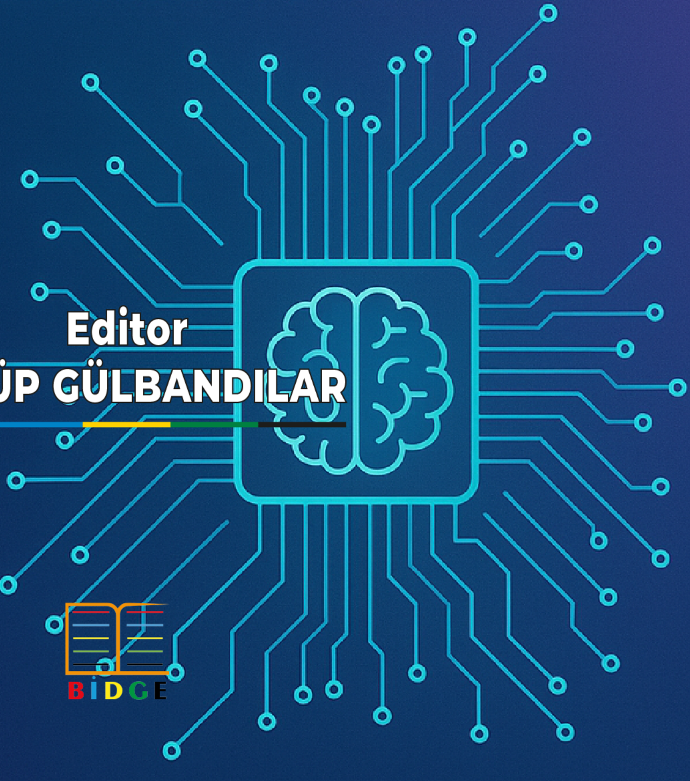
AI-DRIVEN DIGITAL ECOSYSTEMS

FROM GENETICS TO IOT FOR A SECURE AND SUSTAINABLE FUTURE



Editor
EYYÜP GÜLBANDILAR

BİDGE

BİDGE Yayınları

AI-DRIVEN DIGITAL ECOSYSTEMS: FROM GENETICS TO IOT FOR A SECURE AND SUSTAINABLE FUTURE

Editor: EYYÜP GÜLBANDILAR

ISBN: 978-625-372-679-9

1st Edition
Page Layout By: Gözde YÜCEL
Publication Date: 2025-06-25
BİDGE Yayınları

Certificate No: 71374
All rights reserved © BİDGE Yayınları
www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

PREFACE

As we stand at the threshold of the digital era, artificial intelligence (AI) has emerged not merely as a tool of technological transformation but as a central force reshaping diverse domains—from life sciences to cybersecurity. This book aims to provide a comprehensive perspective on AI's role, beginning with its impact on genetic systems and extending to intelligent processing in complex structures, as well as the sustainability and security of next-generation IoT ecosystems.

In today's world, genetic systems, when combined with big data, are revolutionizing fields such as disease diagnosis and treatment planning. AI plays a pivotal role in interpreting biological data and developing clinical decision support systems. Meanwhile, multilayer segmentation techniques in imaging technologies are vital for accurately identifying regions of interest, particularly in medical analysis and biological structures.

Next-generation Internet of Things (IoT) solutions go beyond smart devices—they are foundational to building secure and sustainable digital infrastructures. The effective operation of these systems depends on advanced security mechanisms and energy-efficient designs. Therefore, cybersecurity is a core focus of this book, discussed not only in terms of technical defense but also in light of ethical and sustainability considerations.

This book is intended as a reference for researchers, engineers, students, and all stakeholders interested in navigating the complexities of our digital future. While maintaining scientific integrity, it strives to offer an interdisciplinary perspective and encourages readers to think critically about the technologies shaping tomorrow.

We hope this work serves as a source of inspiration for all those who seek to understand and contribute to an AI-powered future.

Dr. Eyyup GULBANDILAR

Editor

İÇİNDEKİLER

CHAPTER 1

TURKİYE'S CYBER SECURİTY STRATEGİES: LEGAL, TECHNOLOGİCAL, AND GLOBAL COMPLİANCE FRAMEWORK

MEHMET ALI TEKELİ[1]
FATIH BAŞÇİFTÇİ[2]

INTRODUCTION

Since the beginning of the 21st century, technology and digitalization have affected all aspects of modern life; It has caused radical changes, especially in economic, social and political fields. However, this rapid technological development has also brought with it serious cyber security threats.

Dynamics such as technological advances on a global scale, the spread of information and digital systems, and the intensive and continuous use of the internet make the sensitivity to the secure protection of information more visible day by day (Aslay, 2017).

Today, cyberattacks have become a national security priority in a wide range of areas, from the privacy of individuals to the integrity of critical infrastructures. This situation has made it necessary for countries to develop policies that increase not only their defense capacities but also their resilience against cyber attacks.

[1] Selçuk Üniversitesi, Bilişim Teknolojileri Mühendisliği Y.L, Orcid: 0000-0002-9080-7929

[2] Prof.Dr., Selçuk Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği, Orcid: 0000-0003-1679-7416

With its strategic geographical location, rapidly developing digital infrastructure, and growing economy, Turkiye is in a vulnerable position against cyber threats. The digitalization of critical sectors such as energy, transportation, health, and finance in the country has made it necessary to develop national strategies to ensure the security of these infrastructures. With the awareness of this need, Turkiye has developed policies covering both the public and private sectors by preparing national cyber security strategies and action plans since 2013.

In this study, Turkiye's national cyber security strategies published in 2013, 2016 and 2020 and the 2024 Cyber Security Action Plan will be examined comparatively. These documents are important resources for understanding how the country's priorities in cybersecurity, policy instruments, and international collaborations are changing.

The main purpose of the study is to analyze the development of Turkiye's cyber security policies over time by revealing the strengths and weaknesses of these documents.

Considering the importance of cyber security and the necessity of effective implementation of national strategies in this field, this study aims to make both academic and practical contributions. In this regard, the study will not only analyze the current situation but also provide concrete recommendations for future strategies.

MATERIAL AND METOT

In this study, Turkiye's National Cyber Security Strategies published in 2013, 2016, 2020 and the 2024 Cyber Security Action Plan will be examined comparatively. The aim of the study is to understand the basic characteristics of the strategies, to determine the strengths and weaknesses of these strategies, and to reveal the evolution of Turkiye's national cyber security policies over time. It will be made in line with the comparison criteria shown in Figure 1 and both qualitative and quantitative analysis methods will be used.

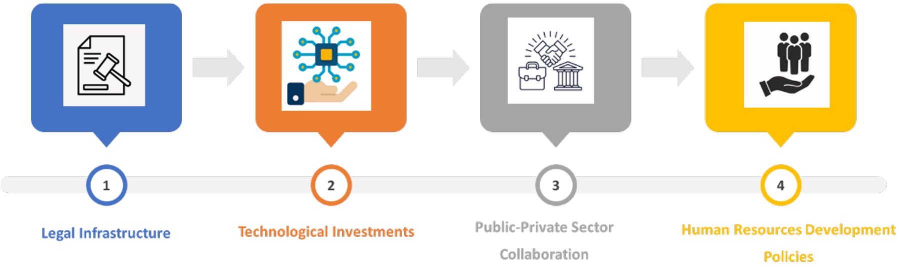

1
Legal Infrastructure
2
Technological Investments
3
Public-Private Sector Collaboration
4
Human Resources Development Policies

Figure 1. Benchmarks

Research Criteria

Turkiye's national strategies in 2013, 2016, 2020 and 2024 Action Plan documents; Legislative regulations will be evaluated within the scope of legal infrastructure, solutions for the protection of critical infrastructures, policies that encourage local software and hardware production, technological investments, joint projects and exercises between the public and private sectors within the scope of public-private sector cooperation, training programs and awareness campaigns within the scope of human resources development policies. Table 1 shows the main headings and explanations used while analyzing the strategy documents.

Table 1. Research Criteria

Criteria	**Explanation**
Legal Infrastructure	• The scope and clarity of the legislative arrangements in the strategies, • Legal regulations to combat cybercrime, • National and international legal compliance.
Technological Investments	• Technological solutions for the protection of critical infrastructures, • Investments in innovative technologies (e.g., artificial intelligence, big data), • Policies that encourage local software and hardware production.
Public-Private Partnership	• Joint projects between the public and private sectors, • Data sharing and coordination mechanisms, • Joint exercises and simulations.
Human Resources Development Policies	• Training programs and awareness campaigns, • Collaborations with universities and certification programs, • Goals of training cyber security specialists.

Comparative Analysis Method

Using the Document Analysis method, Turkiye's national strategies in 2013, 2016 and 2020 and the 2024 Action Plan will be examined through official documents. The objectives, implementation methods and performance indicators specified in the documents will be analyzed.

The research was based on the comparative analysis method. Strategies and action plans were compared in the light of the following criteria:

1. Criterion-Based Comparison: Within the framework of the four main headings mentioned above (legal infrastructure, technological investments, public-private partnership, human resources development), each strategy and action plan will be compared in detail. In this context, the strengths and weaknesses of each period will be identified and emphasis will be placed on changes in plans.
2. Inclusion of an International Perspective: Comparisons will be made with similar strategies of other countries to evaluate the compliance of Turkiye's cyber security strategies with international standards. This benchmark will provide a context for understanding Turkiye's international position in the field of cybersecurity.
3. Historical Analysis: The transformation of Turkiye's national strategies over time is examined and the economic, political and technological contexts of this transformation are discussed.

Data Sources and Collection Method

The scope of the study is limited to the evolution of policies in Turkiye's national cyber security strategies and the position of these policies in the international context. The materials used were derived from the following primary sources:

Official Documents: Turkiye's national cyber security strategy and action plan documents, Academic Publications: Research articles on cyber security strategies,

International Reports: OECD (Organisation for Economic Co-operation and Development), NATO and other international organizations on cyber security.

Since the study was based solely on official documents, there was limited access to practical information such as implementation processes and one-on-one feedback from the field. In addition, limitations in international data have limited in-depth analysis in certain areas of the benchmarking process.

Care has been taken to ensure that the documents used in this study are taken from up-to-date, official and reliable sources. The comparison was carried out based on objective criteria and the social, economic and technological contexts of different periods were taken into account in the analysis process.

SEPARATE REVIEW OF TURKIYE'S STRATEGY AND ACTION PLANS

In this section; The action plans published in 2013, 2016, 2020 and 2024, which are the most important legal infrastructure in the field of cyber security in our country, will be examined in detail within the framework of legal infrastructure, technological investments, public-private sector cooperation, and human resources development.

2013 National Cybersecurity Strategy

The 2013 strategy, Turkiye's first national cyber security strategy, has been a critical starting point for raising awareness and coordination on cyber security. This document has provided an important basis for shaping the strategic approach of both public and private sectors to cyber threats (National Cybersecurity Strategy 2013-2014, 2013). It is Turkiye's first comprehensive strategic document in the field of cyber security.

1. Legal Infrastructure: The 2013 National Cyber Security Strategy is the first document in which Turkiye established a legal framework for cyber security. With this strategy, the establishment of the Cyber Security

Board is envisaged and the role of this board in directing the implementation of national strategies is specified. The Cyber Security Board is designed as a central coordination mechanism to combat cyber threats (National Cybersecurity Strategy 2013-2014, 2013). However, it has been observed that the general scope of legal regulations is limited and more specific regulations are needed.

2. Technological Investments: Regarding technological investments, the 2013 strategy provided general targets for the protection of critical infrastructure. However, no concrete roadmap has been determined for technological capacity building and producing domestic solutions. The strategy is focused more on awareness-raising and core capacity-building efforts.
3. Public-Private Partnership: Public-private partnership was addressed in the 2013 strategy but remained in a very limited scope. The private sector was encouraged to play a supporting role, but it was not stated that this cooperation should be strengthened through sustainable mechanisms (National Cyber Security Strategy 2013-2014, 2013).
4. Human Resources: Training and awareness-raising activities were among the key elements of the 2013 strategy. However, no clear target or implementation plan has been presented for the training of qualified specialists. The focus is more on raising public awareness and sharing information at a basic level (2013-2014 National Cyber Security Strategy, 2013).

2016 National Cybersecurity Strategy

The 2016 strategy aimed to address the shortcomings of the 2013 strategy and to further Turkiye's cyber security policies. This strategy has adopted a more comprehensive approach, with elements such as the development of domestic technologies and the strengthening of national independence.

1. Regulatory Infrastructure: The 2016 strategy expanded legal regulations and focused on more specific regulations. In particular, important legal reforms such as the Law on the Protection of Personal Data (KVKK) were implemented in this period. With this strategy, the first study was carried out to integrate international cooperation mechanisms into the legal infrastructure.
2. Technological Investments: In terms of technological investments, the 2016 strategy highlighted policies that encourage domestic software and hardware production. More concrete targets have been set for the development of security technologies for critical infrastructures. In addition, it is aimed to improve the technological infrastructure for faster detection and elimination of cyber threats (National Cyber Security Strategy 2016-2019, 2016).
3. Public-Private Partnership: Public-private sector cooperation has been developed with a more integrated model during this strategy period. Joint studies were carried out with the private sector on the protection of critical infrastructures and the establishment of regular cooperation platforms was encouraged.
4. Human Resources: The 2016 strategy has implemented training programs aimed at training qualified specialists. In cooperation with universities, specialization programs in the field of cyber security have been initiated. In addition, awareness campaigns have been implemented extensively to reach a wider audience (National Cyber Security Strategy 2016-2019, 2016).

2020 National Cybersecurity Strategy

The 2020 strategy aimed to ensure the integration of innovative technologies and increase cyber resilience in light of global developments in the field of cybersecurity. Security of critical infrastructures and the development of national digital capacity are the main focuses of this strategy (National Cyber Security and Action Plan 2020-2023, 2020).

1. Legal Infrastructure: The 2020 strategy modernized the legal infrastructure and introduced more effective regulations to combat cybercrime. With international cooperation agreements, the compliance of the legal framework with international standards has been increased.
2. Technological Investments: The integration of innovative technologies such as artificial intelligence, big data analytics, and the Internet of Things into cyber security strategies has gained priority in this period. In addition, investments in technology development projects at the local and national level have been increased (2020-2023 National Cyber Security and Action Plan, 2020).
3. Public-Private Partnership: The 2020 strategy has made public-private sector cooperation more sustainable. Practical applications such as the establishment of data sharing standards and joint exercises have been developed.
4. Human Resources: Collaborations with universities have been further expanded, special certification programs have been implemented, and more concrete targets have been set for the training of young professionals.

2024 Cyber Security Action Plan

The 2024 Cyber Security Action Plan aims to build on previous strategies and set more concrete and feasible targets. The action plan also adopts an integrated cyber security approach to digital transformation processes (2024-2028 National Cyber Security and Action Plan, 2024).

1. Legal Infrastructure: Specific regulations have been introduced for the protection of critical infrastructures,

and international norms have been taken into account to increase legal compliance.

2. Technological Investments: Blockchain-based security solutions, artificial intelligence applications, and supporting national R&D centers are among the key elements of the 2024 action plan. In addition, cybersecurity investments for digital transformation projects have been increased (2024-2028 National Cyber Security and Action Plan, 2024).
3. Public-Private Partnership: The 2020 strategy has made public-private sector cooperation more sustainable. Practical applications such as the establishment of data sharing standards and joint exercises have been developed (National Cyber Security, 2025).
4. Human Resources: Collaborations with universities have been further expanded, special certification programs have been implemented, and more concrete targets have been set for the training of young professionals.

COMPARATIVE ANALYSIS

A comparative analysis of national cybersecurity strategies and the 2024 Cybersecurity Action Plan is critical to understanding how Turkiye's cybersecurity policies have changed and evolved over time. In this section, strategies will be analyzed under four main headings and their strengths, weaknesses, differences and continuity will be discussed.

Legal Infrastructure

The legal infrastructure elements in Turkiye's cyber security strategies have expanded and deepened over time. Below, the action plans published in our country are compared within the framework of the legal infrastructure.

1. 2013 Strategy: With the establishment of the Cyber Security Board, the first steps towards cyber security coordination were taken. However, in this period, the

legal infrastructure was largely at a basic level, and more comprehensive and binding regulations were missing.

2. 2016 Strategy: In this period, the legal framework was expanded and important regulations such as the Personal Data Protection Law (KVKK) were made. In addition, a stronger effort has been made to harmonize national regulations with international law.
3. 2020 Strategy: The legal infrastructure has been modernized to support international cooperation. Innovative approaches and harmonized international norms have been adopted in the fight against cybercrime.
4. 2024 Action Plan: This plan, is aimed to clarify the legal regulations for implementation and the preparation of specific regulations on the protection of critical infrastructures has been brought to the agenda. Improving compliance with international standards has been a key priority of the plan.

Since 2013, the legal infrastructure has transformed from a general framework to a more specific and practice-oriented structure. In the 2024 action plan, the focus of this transformation on concrete implementation processes is considered an important step in terms of eliminating legal deficiencies.

Technological Investments

The strategies and action plan show marked progress in approaches to technological investments. Below, the action plans published in our country are compared within the framework of technological investments.

1. 2013 Strategy: Targets for technological investments remained at a general level, and the development of technological solutions for the protection of critical infrastructures was encouraged. However, the paucity of concrete steps is noteworthy.

2. 2016 Strategy: Concrete policies have been developed to support local software and hardware production. A serious vision has been put forward to increase Turkiye's technological independence.
3. 2020 Strategy: Innovative technologies such as artificial intelligence, big data analytics, and the Internet of Things are integrated into the strategies. This has increased Turkiye's capacity to adapt to global technological developments.
4. 2024 Action Plan: The plan aims to concretize technological investments at an advanced level. Blockchain-based security solutions and the establishment of national R&D centers are integrated into digital transformation projects.

In terms of technological investments, the 2024 Action Plan set out the most concrete and comprehensive targets, while the 2013 strategy remained more general and at the initial level. The emphasis on innovative technologies, especially in the 2020 and 2024 strategies, is an important development in terms of the sustainability of technological progress.

Public-Private Partnership

Public-private partnership has been increasingly strongly emphasized and developed across strategies. Below, a comparison of the action plans published in our country within the framework of public-private sector cooperation has been made.

1. 2013 Strategy: Public-private cooperation was supported, but regular mechanisms to improve this cooperation were not included.
2. 2016 Strategy: More integrated cooperation models have been developed with the private sector with a focus on critical infrastructures. Joint projects and regular dialogue mechanisms have been proposed.
3. 2020 Strategy: New standards have been set for the sustainability of cooperation between the public and

private sectors and data sharing processes have been modernized. Exercises and joint work have been increased.

4. 2024 Action Plan: Public-private sector relations were discussed at a more advanced level, cooperation in innovation projects was emphasized, and the organization of joint exercises was among the concrete objectives.

Public-private sector cooperation has undergone an evolution process since 2013 and has reached the level of institutionalization with the 2024 Action Plan (Kurnaz & Önen, 2019). In particular, the determination of standards for data sharing has made this cooperation more concrete and effective.

Human Resources Development Policies

Human resource development policies have evolved throughout strategies and become more comprehensive. Below, the action plans published in our country are compared within the framework of human resources development policies.

1. 2013 Strategy: Awareness campaigns were organized, but the targets of training qualified experts remained limited.

2. 2016 Strategy: Training programs were initiated in cooperation with universities and public institutions and initiatives were made to train cyber security experts.

3. Strategy **2020**: Dedicated certification programs have encouraged young professionals to specialize in cybersecurity. Awareness campaigns have reached wider audiences.

4. 2024 Action Plan: In order to develop human resources, concrete goals such as disseminating international certificates, opening graduate academic programs and raising awareness at an early age in schools have been put forward.

In terms of human resource development, the 2024 action plan offered the most comprehensive approach. Initiating awareness campaigns at an early age is important for long-term capacity building goals.

FUTURE CYBER SECURITY STUDIES FOR TURKIYE

Turkiye should aim to achieve full integration with GDPR and other global data protection regulations to better align with international standards in its cybersecurity strategies. Within the framework of NATO cooperation, more joint exercises and capacity building projects can be participated. By examining the public-private sector cooperation models of countries such as the USA and Germany, adaptive models should be developed for Turkiye in accordance with local dynamics.

Strengthening Technological Independence

Turkiye's investments in domestic technologies should form the basis of future cyber security strategies. In particular, emphasis should be placed on developing more innovative solutions by reducing foreign dependency. The steps that can be taken in this context are listed below.

1. Artificial Intelligence and Big Data: The use of artificial intelligence algorithms in the detection and prevention of cyber threats should be expanded. With big data analytics, proactive security measures can be taken by learning from past attacks.

2. Quantum Computing: Quantum computing is expected to revolutionize the field of cybersecurity in the future. Turkiye can increase its international competitiveness by investing in research and development (R&D) projects in this field.

3. Blockchain Technology: Blockchain has a critical role in ensuring data security and accuracy. Blockchain solutions should be given more space in public and private sector applications.

Deepening International Cooperation

Turkiye's cooperation with international organizations such as NATO, ITU, OECD and the EU can enable it to gain a stronger cyber security position in the future. In this context, it should be a partner in global exercises and the effectiveness of international data sharing mechanisms should be increased. In this context, the collaborations that can be made in the international arena are listed below.

1. Global Exercises: Turkiye's greater participation in joint global exercises with international organizations such as NATO and ITU will increase its coordination competence in times of crisis.
2. International Data Sharing Mechanisms: Turkiye's leading role in data sharing platforms where it can contribute to global threat analysis will increase its effectiveness in the cyber security community.
3. Full Compliance with EU Cybersecurity Standards: Turkiye can strengthen its international cooperation capacity by fully complying with GDPR and other European Union regulations.

Institutionalization of Public-Private Sector Cooperation

Making the cooperation models between the public and private sectors more effective and efficient is very important in the field of cyber security as well as in the social and economic fields. With the joint investment projects of the public and private sectors, the development of the field of cyber security and the determination of common security standards can be ensured. The studies that can be done in this field are listed below.

1. Co-Investment Platforms: Greater involvement of the private sector in innovation projects can ensure the effective use of financial resources.
2. Standardization: Common standards should be developed for data sharing and critical infrastructure protection. These standards can be harmonized with internationally accepted norms.

3. Joint Exercises: Regular simulations and exercises between the public and private sectors should be encouraged to gain practical experience.

Human Resources and Training Strategies

Turkiye's young population has great potential in the field of cyber security. Training and capacity building strategies should be strengthened to use this potential effectively.

1. Cyber Security Awareness in Schools: Cybersecurity modules should be included in the education curriculum in primary and secondary schools.
2. University and Certification Programs: Cybersecurity departments in universities should be expanded and the accessibility of international certificates should be increased.
3. Professional Development: The professional development of cyber security experts should be supported by organizing continuous training programs in cooperation with the public and private sectors.

Digital Government and Social Security Applications

Digital government and social security applications are systems that aim to facilitate the lives of individuals with the integration of modern technology into public services. In Turkiye, these applications have become accessible in a wide range of applications, especially with the services offered through the e-Government Gateway.

Turkiye can carry out the following studies in this context, aiming to strengthen its digital government infrastructure.

1. Security of E-Government Systems: Artificial intelligence and blockchain technologies can be used to protect e-Government services.
2. Individual Awareness Campaigns: Broader awareness campaigns should be organized to increase cyber security awareness throughout society.

5.6. Long-Term Strategic Priorities

Long-term strategic priorities are needed to strengthen Turkiye's future goals in the field of cyber security and to increase its international competitiveness. These priorities focus on key areas such as digitalization, international cooperation, human resource development and technological independence.

The long-term plans that Turkiye can make in line with its goal of leadership in cyber security are listed below.

1. National Cyber Security Academy: An academy can be established that focuses on expert training and research activities.
2. Global Cybersecurity Leadership: Turkiye can make a wider impact by taking a leadership role in the international cybersecurity community.
3. Cyber Security Funds: Innovation can be encouraged by creating special funds for R&D studies.

COMPARISON OF TURKIYE'S CYBER SECURITY STRATEGIES WITH INTERNATIONAL STANDARDS

Comparing Turkiye's cyber security strategies with international standards is important to understand the country's strengths and areas that need to be improved in this area. In this section, comparisons will be made with NATO, ITU (International Telecommunication Union), USA, Germany, Estonia, India and South Korean countries within the framework of legal infrastructure, technological investments, international cooperation, public-private sector cooperation, human resources of the cyber security strategies published by our country. The countries to be compared are selected from countries that work in the field of cyber security in the world or are exposed to attacks.

Compliance of Turkiye's Cyber Security Strategies with International Standards

Turkiye's cyber security strategies aim to increase international cooperation in the fight against global cyber threats. In this context, Turkiye's cyber security policies strive to comply with various international norms and standards. The factors that determine international standards in the field of cyber security are listed below.

1. NATO Cyber Security Framework: Turkiye, as a NATO member, has adapted to NATO's cyber security strategies. In line with NATO's common cyber defense policies, Turkiye has developed policies to protect critical infrastructures, early detection of cyber threats, and increase coordination (MIL, et al., 2014).
2. ITU (International Telecommunication Union) Global Cyber Security Index (GCI): ITU is a specialized organization affiliated with the United Nations. ITU operates in the field of global telecommunications and information communication technologies to set standards, encourage international cooperation and support technological development.

ITU's Global Cybersecurity Index (GCI) is a measurement tool that evaluates countries' cybersecurity capabilities. This index enables countries to be ranked according to criteria such as legal infrastructure, technical capacity, organizational structure, capacity building and international cooperation. Through this index, the ITU helps countries identify their strengths and weaknesses in the field of cybersecurity and promotes the development of cybersecurity policies at the global level. In Table 2, Turkiye's cyber security index by years according to ITU is given.

*Table **2.** Turkiye's GCI Index according to ITU (Global-Cybersecurity-Index, 2025)*

Year	GCI Score	World Rankings	European Rankings	Evaluation Note
2017	0.683	43	17	Strengthening the basic infrastructure
2018	0.712	39	15	New regulations have been put in place
2020	0.789	33	12	Critical infrastructure protection increased
2022	0.821	29	10	Strategy and technology investment is evolving
2024	0.853	25	8	International cooperation has improved

Turkiye's GCI ranking shows progress in legal infrastructure, technical capacity, organizational structure, international cooperation and capacity building efforts. The increase in Turkiye's rankings and scores in recent years supports the strengthening of internationally harmonized policies.

1. European Union Cybersecurity Norms: Turkiye has taken significant steps towards compliance with the EU's General Data Protection Regulation (GDPR). In particular, KVKK can be considered a local reflection of GDPR. This alignment demonstrates Turkiye's approach to European standards in data protection policies.

Comparative Evaluation Between Turkiye and Other Countries

The compliance of Turkiye's cyber security policies with international standards shows some commonalities and differences compared to other countries. The cyber security policies of our country have been compared with the USA, Germany, Estonia, India and South Korea countries, which carry out the most legal and technical studies in this field.

Compared to countries with advanced cyber security infrastructures such as the USA and Estonia, Turkiye emphasizes domestic independence, especially in terms of technological investments. This situation provides Turkiye with a strong strategic advantage in reducing foreign dependency.

Compared to European Union countries, Turkiye's efforts to comply with GDPR and regulations such as KVKK show that it is advancing in the field of data protection policies. However, Turkiye needs to make more progress in public-private coordination and cooperation against cyber threats.

Compared to Estonia, Turkiye's focus on innovative technologies and digital transformation projects is noteworthy. However, Estonia's digital government models and rapid response capacity can be a source of inspiration for Turkiye.

Table 3 shows the comparison of Turkiye's cyber security policies with the USA, Germany, Estonia, India and South Korea according to the criteria of legal infrastructure, technological investment, public-private sector cooperation, and human resources.

Table 3. Comparison of Turkiye's Cyber Security Policies

Criteria	**Turkiye**	**US**	**Germany**	**Estonia**	**India**	**South Korea**
Legal Infrastructure	KVKK, critical infrastructure protection regulations, NATO compliance.	Comprehensive cybercrime laws at the federal level, policies that overlap with GDPR.	GDPR compliant policies, and industry regulations.	Fast, simple cybersecurity laws.	The Data Protection Act (2019) is one of the national cybersecurity strategies.	National Cyber Security Law and detailed regulations.
Technological Investments	Focused on domestic production, artificial intelligence, blockchain and critical infrastructure.	Large-scale artificial intelligence projects, government-sponsored technology initiatives.	Investments in industrial infrastructure and innovative technology.	Digital infrastructure defense systems, e-Government integration.	Digital infrastructures supporting domestic production and Start-Up India initiative.	Serious investments in 5G technology, artificial intelligence and IoT infrastructures.
International Cooperation	Cooperation within the framework of NATO and ITU, OECD harmonization.	Global information sharing and intelligence collaboration with the Five Eyes Alliance.	NATO and EU integration, information sharing and coordination mechanisms.	Full integration into NATO, regular participation in cyber exercises.	Coordination with BRICS countries, global digital infrastructure cooperation.	Joint initiatives and information sharing with ASEAN, ITU and NATO.
Public-Private Partnership	It is being made more systematic with the 2024 targets.	Collaboration mechanisms with technology companies are widely implemented.	Large-scale public-private partnership with academic and sectoral focus.	Fast, flexible and innovation-oriented collaboration.	Partnerships with the private sector and digital initiatives are supported in infrastructure investments.	Strong collaborations with technology companies in critical infrastructure projects.
Human Resources	University collaborations, international certification programs.	Large-scale federal funding projects for cybersecurity professionals.	Training and industry-oriented certification programs.	Training and digital skill development projects, and expert training were accelerated.	Broad-based technical education and employment policies.	Digital economy and manpower development programs, extensive awareness campaigns.

6.3. The International Position of Turkiye's Cyber Security Policies

Turkiye is developing its cyber security policies on a global scale and strengthening international cooperation. The studies carried out in this context are listed below.

1. NATO Exercises: Joint exercises with NATO increase Turkiye's international cyber security capacity. These exercises support the sharing of experience and the flow of information for the protection of critical infrastructures.
2. ITU Global Cyber Security Index: Turkiye's rise in the index rankings shows the progress made in the field of international cooperation and national capacity building.
3. International Cooperation Mechanisms: Turkiye's participation in international mechanisms for data sharing and joint threat analysis can be considered an important step in the fight against global cyber threats.

Future Perspectives and Recommendations for Turkiye

The following suggestions are presented for Turkiye to increase its international compliance and become an effective cyber security player at the global level.

1. Increasing International Exercises: Turkiye's participation in joint exercises with NATO, ITU and other global organizations should be further increased.
2. Strengthening Data Sharing Mechanisms: It can be ensured that Turkiye plays a leading role in global data sharing platforms.
3. Integration of Public-Private Partnership with International Models: The public-private partnership models of countries such as the USA and Estonia can be adapted to Turkiye.

4. Dissemination of Digital Government Applications: Inspired by Estonia's digital government models, the scope of digital services in Turkiye can be expanded.

Turkiye's cyber security strategies have gradually become comprehensive since 2013. While the 2013 strategy provided a basic framework, the 2016 and 2020 strategies brought more specific and innovative approaches. The 2024 Action Plan, on the other hand, has taken all these achievements to an advanced stage by combining them with concrete implementation targets.

RESULTS AND RECOMMENDATIONS

Results

Turkiye's national cyber security policies have developed since 2013 and have taken on a more comprehensive structure over time. Below are the key takeaways from the study.

1. Legal Infrastructure: The 2013 strategy enabled Turkiye to establish an institutional and legal foundation for cybersecurity. However, these regulations were limited to a general framework. The 2016 strategy expanded the legal regulations and implemented critical regulations such as the Personal Data Protection Law (KVKK). The 2020 and 2024 processes, on the other hand, have developed regulations that are more modern and compatible with international standards. Especially with the 2024 Cyber Security Action Plan, the legal infrastructure has turned to more concrete targets and an application-oriented structure.
2. Technological Investments: Technological investments have continuously gained importance since the strategies in 2013. The 2016 strategy increased the emphasis on domestic production and aimed to strengthen national capacity. The 2020 strategy strategically incorporated innovative technologies such as artificial intelligence and big data analytics. The 2024 action plan, on the other hand, takes these gains further and sets specific

technological goals such as blockchain, digital transformation and R&D investments.

3. Public-Private Partnership: Cooperation between the public and private sectors has been progressively strengthened throughout the strategies. While cooperation was at the initial level in the 2013 strategy, these relations were supported by more integrated models in 2016 and beyond. In particular, the 2024 action plan envisaged concrete cooperation mechanisms such as innovation projects and data sharing standards.
4. Human Resources: Human resources development policies, which were mostly handled at the level of awareness in 2013, have deepened with training and certification programs in the 2016 and 2020 strategies. The 2024 action plan, on the other hand, has set out wide-ranging goals such as international certification, academic programs and early age awareness campaigns.

In Table 4, the evaluation of the action plans prepared by our country in line with the criteria of legal infrastructure, technological investments, public-private partnership and human resources is given.

Table 4. Comparison of Action Plans

Criteria	2013 Strategy	2016 Strategy	2020 Strategy	2024 Action Plan
Legal Infrastructure	Establishment of the Cyber Security Board; Establishment of the general legal framework.	Entry into force of KVKK; international legal harmonization efforts.	Innovative regulations in the fight against cybercrime; international agreements.	Specific regulations for critical infrastructures; Practice-oriented legislative developments.
Technological Investments	General objectives for the protection of critical infrastructures.	Domestic production-oriented policies; emphasis on technological independence.	Integration of innovative technologies such as artificial intelligence and big data into strategies.	Investments focused on blockchain-based security, R&D centers, and digital transformation.
Public-Private Partnership	Public-private cooperation is at the initial level; limited projects.	Increased integration on critical infrastructures; mechanisms for regular dialogue.	Development of data sharing standards; Sustainable cooperation models.	Exercises, innovation projects and implementation of common data security standards.
Human Resources	Awareness campaigns; general educational work.	Initiation of training programs; and collaborations with universities.	Certification programs; Supporting young professionals.	International certificates; postgraduate training, early years awareness campaigns.

Turkiye's cyber security strategies have evolved from documents that initially drew more general frameworks to comprehensive policies that focused on concrete practices and advanced technological targets over time. This process demonstrates Turkiye's determination to comply with international cybersecurity standards and strengthen its national capacity. However, it is seen that sustainability should be ensured in implementation and audit processes.

Recommendatıons

Various suggestions can be made to make Turkiye's cyber security strategies more effective and to increase its competitiveness at the international level. These recommendations cover improvement and development strategies in key areas such as technology, international cooperation, human resources and legal regulations.

The following suggestions are presented to further strengthen Turkiye's cyber security strategies.

Strengthening Implementation and Supervision Mechanisms:

- It is necessary to ensure the transformation of strategies into concrete targets and to establish independent audit mechanisms to monitor the applicability of these goals.
- It is recommended to implement regular performance evaluation processes for the public and private sectors.

Increasing Technological Independence:

- R&D investments for the development of domestic and national technologies should be increased.
- Domestic solutions should be developed to reduce foreign dependency on critical infrastructures.
- Long-term investment plans should be prepared for advanced technologies such as artificial intelligence, quantum computing, and blockchain.

Deepening International Cooperation:

- Exercises and Simulations: Turkiye should regularly participate in cyber security exercises with organizations such as NATO and ITU and integrate the knowledge and experience gained from these exercises into its strategies.
- Information Sharing and Coordination: Leadership in cyber threat analysis should be assumed by taking an active role in international data sharing platforms.

- Compliance **with Standards:** Full integration with the European Union's data protection regulations such as GDPR should be ensured and compliance with other global standards should be increased.

Training and Human Resource Development:

- Starting from primary school, cybersecurity-related modules should be added to educational curricula to create awareness of cybersecurity.
- Departments and graduate programs for specialization in the field of cyber security should be increased in universities.
- With the cooperation of the public and private sectors, the accessibility of international certificates should be increased and professional development should be supported.

Institutionalization of Public-Private Partnership Models:

- Common platforms should be established to ensure regular coordination between the public and private sectors.
- Common standards should be applied in the processes of data sharing and protection of critical infrastructures.

Dissemination of Cyber Security Awareness to the Society:

- Broader campaigns should be organized to raise awareness of cybersecurity across society.
- In particular, the cybersecurity knowledge of individuals and small businesses should be increased.

REFERENCES

National Cybersecurity Strategy 2016-2019. (2016). hgm.uab.gov.tr: Retrieved from https://hgm.uab.gov.tr/uploads/pages/siber-guvenlik/2016-2019guvenlik.pdf [Access Date: 04.01.2023]

2020-2023 National Cyber Security and Action Plan. (2020). T.R. Ministry of Transport and Infrastructure: Retrieved from https://hgm.uab.gov.tr/uploads/pages/siber-guvenlik/ulusal-siber-guvenlik-stratejisi-ep-2020-2023.pdf [Access Date: 10.12.2023]

2024-2028 National Cyber Security and Action Plan. (n.d.). Retrieved from https://www.uab.gov.tr/: https://www.uab.gov.tr/uploads/pages/siber-guvenligin-yol-haritasi-yerli-ve-milli-tekno/ulusal-siber-guvenlik-stratejisi-2024-2028.pdf [Access Date: 20.09.2024]

ASLAY, F. (2017). Cyber Attack Methods and Turkiye's Cyber Security Current Situation Analysis. *International Journal of Multidisciplinary Studies and Innovative Technologies, 1(1),* pp. 24-28.

Global-Cybersecurity-Index. (2025). https://www.itu.int/: Retrieved from https://www.itu.int/en/ITU-D/Cybersecurity/Pages/Global-Cybersecurity-Index.aspx

Kurnaz, S., & Önen, S. (2019). Turkiye's cyber security strategies in the process of harmonization with the European Union. *Ankara University Faculty of Law Journal*, p. 68(4), 123-140. https://dergipark.org.tr/en/download/article-file/819421 [Access date: 24.03.2025]. Retrieved from

MİL, H. İ., Gözübenli, M., Harmancı, F. M., & Cevdet, Z. (2014). Turkiye's cyber security strategies. *Journal of Law and Informatics, 5(2), 45-67.*, pp. 5(2),45-46. https://www.researchgate.net/publication/349052551_TURKIYE'NIN_SIBER_GUVENLIK_STRATEJILERI_2014_Mil_HI_Gulep_S_ve_Unal_A [Access date: 24.03.2025]. Retrieved from

National Cybersecurity. (2025). https://cyberartspro.com/: https://cyberartspro.com/ulusal-siber-guvenlik-stratejisi-ve-2024-2028-eylem-plani/ [Access Date: 08.03.2025]. Retrieved from

CHAPTER 2

EMERGING TECHNOLOGIES AND SOLUTIONS FOR NEXT-GENERATION IOT ECOSYSTEMS

ÖZLEM BATUR DİNLER[1]

Introduction

Once a futuristic vision, the Internet of Things (IoT) has rapidly evolved into a pervasive and transformative force across virtually every sector. This chapter traces the foundational journey of IoT ecosystems, highlighting the technological and architectural shifts that have given rise to next-generation implementations.

In its early stages, IoT was characterized by simple applications such as environmental monitoring, asset tracking, and basic automation. These first-generation systems relied on rudimentary sensors that collected data and transmitted it to centralized cloud infrastructures for storage and analysis. Decision-making remained largely manual, with human operators interpreting the data and initiating responses. Intelligence in these systems was concentrated in the cloud, while edge devices functioned merely as passive data collectors.

[1] Assistant Professor, Siirt University, Department of Computer Engineering, 0000-0002-2955-6761

Although groundbreaking, this centralized model revealed significant limitations as IoT deployments scaled. Latency issues hindered time-sensitive applications, and the explosion of data volumes strained available bandwidth. Security and privacy became critical concerns as sensitive information traveled across distributed networks. Moreover, the reliance on uninterrupted connectivity exposed vulnerabilities in remote or intermittently connected environments.

Next-generation IoT ecosystems have embraced a decentralized, intelligent approach to address these challenges. Rather than relying on rigid hierarchies, modern IoT architectures distribute computing power and decision-making capabilities across the network. This shift makes systems more responsive, resilient, and adaptive to dynamic operational demands.

This chapter explores the core technological foundations that underpin this evolution, focusing on three key paradigms: the edge-to-cloud continuum, cognitive IoT systems, and digital twin integration. Together, these paradigms redefine the capabilities of IoT—from merely sensing and monitoring to analyzing, predicting, and autonomously responding to complex real-world conditions. We also examine advances in connectivity technologies that serve as the backbone for these intelligent, scalable ecosystems.

By understanding these foundational components, we set the stage for the following chapters to delve into specific technologies, application areas, and the future trajectory of next-generation IoT ecosystems.

The Evolving IoT Landscape: From Connected Devices to Intelligent Ecosystems

The transformation of IoT from concept to reality marks a pivotal shift in technological evolution over the past decade. Early IoT deployments were relatively simple—sensors collected basic environmental or operational data and transmitted it to centralized cloud platforms for rudimentary analysis. These systems focused on monitoring, leaving decision-making to human operators based on the collected insights. The architecture was predominantly hierarchical and centralized, with intelligence concentrated in cloud systems rather than distributed across the network.

Today's next-generation IoT ecosystems have evolved dramatically. Intelligence is now distributed throughout the network, from sophisticated edge devices to intermediate fog nodes and cloud platforms. This shift enables autonomous decision-making at multiple levels, reducing latency and minimizing dependence on centralized systems. The focus has moved beyond simple data collection to complex event processing, predictive analytics, and autonomous intervention. Early IoT systems were primarily reactive, whereas modern implementations are increasingly proactive—anticipating issues and adapting to changing conditions with minimal human involvement.

Figure 1 illustrates the early IoT architecture, which relied on a centralized structure. In this model, simple sensors collect essential data and transmit it directly to a central cloud platform, where all intelligence is concentrated. The diagram shows multiple sensors at the bottom connecting to a single cloud system at the top, reflecting the hierarchical nature of early IoT deployments. This architecture primarily focused on basic monitoring, with human operators interpreting data and making decisions accordingly.

Figure 1: *Centralized IoT Architecture: Early-Stage Model.*

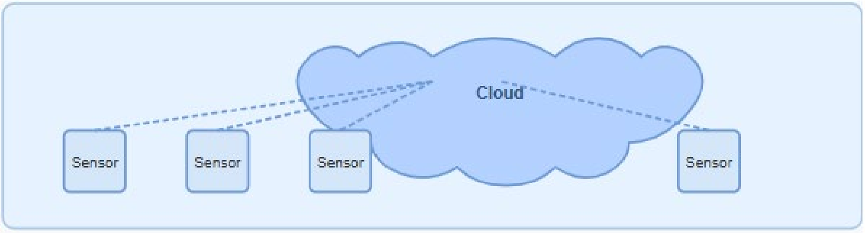
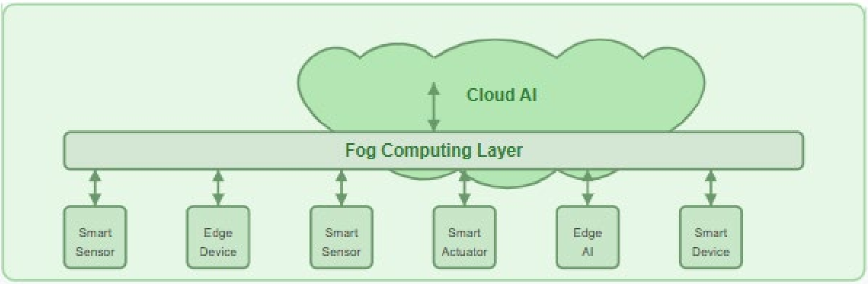


Figure 2 depicts the transition to modern IoT architectures, where intelligence is distributed across the network. Unlike early models, this system integrates smart edge devices, a fog computing layer, and AI-powered cloud platforms. Bidirectional arrows signify autonomous decision-making across multiple layers, ensuring lower latency and reduced reliance on centralized systems. Smart sensors, edge devices, actuators, and AI-driven analytics work together to enable complex event processing, predictive insights, and real-time autonomous responses.

Figure 2: *Distributed Intelligence in Modern IoT Systems*



This shift is driven by three key paradigms:

Edge-to-Cloud Continuums

Edge-to-cloud continuums represent a fundamental shift in how IoT allocates and utilizes computational resources. Rather than treating edge and cloud as separate domains, modern architectures implement a dynamic continuum where processing occurs at the optimal location based on latency requirements, bandwidth availability, power constraints, and data sensitivity.

This approach allows computational tasks to migrate across the continuum based on evolving conditions. For example, an autonomous vehicle processes critical safety-related data locally while offloading non-urgent analytics to the cloud. Similarly, a manufacturing system performs real-time quality control at the edge while leveraging cloud resources for long-term process optimization.

Cognitive IoT Systems

Cognitive IoT systems integrate artificial intelligence at multiple levels, from sensor-level signal processing to edge analytics and cloud-based deep learning. This enables previously unattainable capabilities, such as natural language understanding for voice-controlled devices, computer vision for automated inspections, and reinforcement learning for autonomous process optimization.

The line between IoT and AI is increasingly blurred, with machine learning models becoming integral to IoT architecture rather than merely consumers of IoT-generated data. Federated learning further strengthens this integration, enabling distributed devices to train models without centralizing sensitive data.

Digital Twin Integration

Digital twins—virtual replicas of physical assets, processes, or systems (Anand, 2024)—bridge the gap between the physical and digital worlds (Rani et al., 2024) , enabling simulation, optimization, and predictive maintenance.

Unlike static models, advanced digital twins continuously (Rani et al., 2024) evolve based on real-time IoT sensor data. This enables **"what-if"** scenario testing, predictive maintenance, and optimization that would be impractical or prohibitively expensive in physical environments. For example, a manufacturing facility can simulate different production schedules before implementing them, while a smart building can optimize HVAC operations based on occupancy and weather conditions.

The convergence of these three paradigms—edge-to-cloud continuums, cognitive IoT systems, and digital twin integration—is creating IoT ecosystems of unprecedented intelligence and adaptability. These systems can sense, analyze, predict, and respond autonomously, unlocking new business models and efficiencies across industries. However, realizing their full potential requires overcoming critical challenges in connectivity, edge intelligence, security, data management, and sustainability—key aspects explored in the following sections.

Connectivity Innovations for Scalable IoT

As the Internet of Things (IoT) scales from thousands to billions of connected devices, the challenges of efficient connectivity and power management grow increasingly complex. To address these challenges, emerging IoT technologies integrate innovations in energy harvesting, ultra-low-power communication, and intelligent network management. These advancements collectively create a seamless and scalable IoT ecosystem capable of operating in diverse environments with minimal energy consumption.

Ambient IoT and Zero-Power Devices

A significant challenge in achieving ubiquitous IoT deployment is the high power consumption of connected devices. Traditional IoT sensors rely on batteries that must be periodically replaced or recharged, leading to substantial maintenance costs and limiting deployment options and device form factors. Ambient IoT technologies offer a groundbreaking solution by enabling devices to operate with minimal or even zero external power sources.

Radio Frequency Energy Harvesting

Radio Frequency (RF) energy harvesting technologies enable IoT devices to extract power from existing electromagnetic fields in the environment. These systems typically employ specialized antennas and efficient power conversion circuits to capture energy from ambient sources such as Wi-Fi (Chen et al., 2023), cellular networks, and television or radio broadcasts. Recent advancements in ultra-low-power rectifiers and power management integrated circuits (PMICs) have significantly improved conversion efficiency, allowing IoT devices to harvest meaningful amounts of energy even from weak RF fields. For example, researchers at the University of Washington and MIT have demonstrated sensors powered by ambient television signals or Wi-Fi transmissions that support basic sensing and communication functions (Liu et al., 2013; Wang et al., 2014). These techniques are particularly effective in urban environments, where RF energy density is relatively high.

Backscatter Communication

Backscatter communication provides an alternative solution to power constraints in IoT devices. Instead of generating their own radio signals—which require substantial power—backscatter devices modulate and reflect existing RF signals to transmit data. This reduces power consumption, enabling communication within microwatt-level power budgets compared to the milliwatts required by conventional radio transmitters. Advanced backscatter systems can achieve data rates of hundreds of kilobits per second while consuming less than 10 microwatts of power, making them suitable for many sensing applications. When combined with energy harvesting techniques, backscatter communication enables fully battery-free wireless sensing systems that can operate indefinitely in suitable environments.

Passive Sensing Technologies

Passive sensing technologies push energy efficiency to the next level by eliminating the need for powered electronics in certain applications. Instead of relying on traditional active components, these systems use materials whose properties naturally change in response to environmental conditions, allowing sensing to occur without an active power source. This innovation enables sensors to function through physical or chemical interactions with their surroundings.

For example, passive sensors can change their resonant frequency in response to changes in temperature, humidity, or the presence of specific chemicals. These sensors can be wirelessly interrogated using external reader devices, much like radio-frequency identification (RFID) tags, but with the added benefit of environmental sensing. Since these sensors require no batteries or active electronics, they can be made very small, lightweight, and inexpensive. Their durability and low maintenance requirements make them ideal for deployment in environments where conventional electronic sensors would be impractical—such as remote locations, harsh industrial settings, or embedded applications requiring long-term reliability.

By integrating passive sensing with other advancements like energy harvesting and low-power communication, IoT systems can achieve new levels of efficiency and scalability. These technologies enable the creation of vast sensor networks that operate indefinitely, requiring minimal human intervention while continuously providing reliable data.

Connectivity Orchestration

As IoT deployments scale from thousands to billions of connected endpoints, managing connectivity becomes increasingly complex. Modern IoT systems integrate multiple wireless technologies, each with unique characteristics and trade-offs, requiring advanced strategies for efficient connectivity management. Next-generation IoT systems adopt connectivity orchestration strategies that dynamically select and transition between communication technologies based on application requirements, network conditions, and energy considerations.

Multi-Radio Access Technology (Multi-RAT) Management

Multi-Radio Access Technology (Multi-RAT) management platforms offer a unified approach to managing diverse wireless technologies within IoT deployments. These systems abstract the complexities of individual wireless protocols, providing a consistent interface to applications while dynamically selecting the optimal communication technology based on contextual factors. For example, a mobile IoT device might use cellular connectivity outdoors, switch to Wi-Fi when entering a building, and switch to Bluetooth Low Energy as the battery level drops. This dynamic selection process considers signal quality, bandwidth requirements, latency, power consumption, and connectivity costs. Advanced Multi-RAT systems incorporate machine learning algorithms that adapt selection strategies based on historical performance, predicted network conditions, and specific application requirements.

Dynamic Spectrum Access

Dynamic Spectrum Access technologies enhance Multi-RAT management by enabling more efficient use of available radio spectrum. Traditional wireless systems operate in fixed frequency bands, often leading to inefficient spectrum utilization. Software-defined radio technologies allow IoT devices to identify and use available spectrum across licensed, unlicensed, and shared bands. Dynamic spectrum access enables IoT deployments to tap into previously underutilized spectrum resources when combined with regulatory frameworks like TV White Spaces and Citizens Broadband Radio Service (CBRS). These technologies are especially valuable in rural or underserved areas with limited connectivity options, as well as in dense urban environments where conventional spectrum bands are congested.

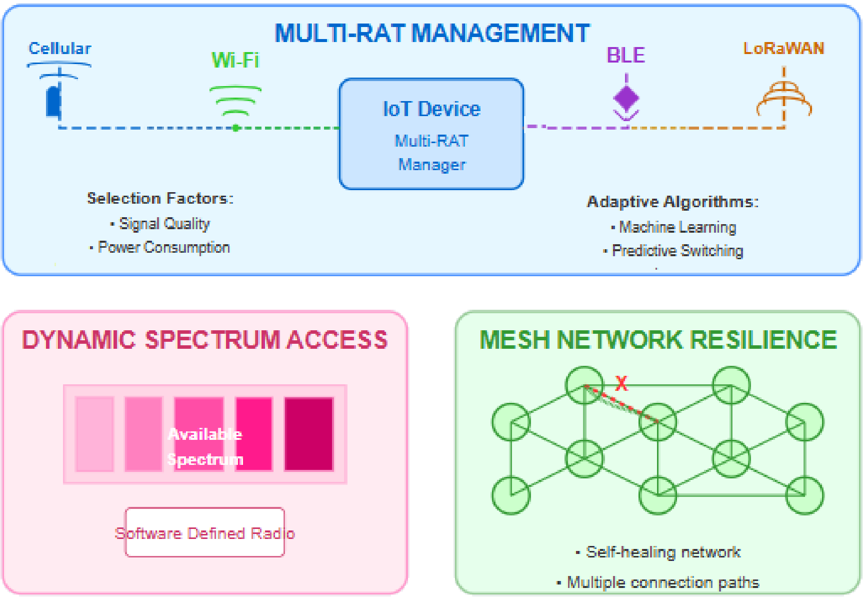
Mesh Network Resilience

Mesh networking represents the third pillar of connectivity orchestration in next-generation IoT deployments. In traditional star-topology networks, where devices connect directly to a central access point, the system can be vulnerable to single points of failure and limited coverage in complex environments. Self-organizing mesh topologies address these limitations by enabling devices to relay messages through multiple paths, dynamically adapting to changing network conditions. Advanced mesh protocols incorporate cognitive routing capabilities that optimize paths based on signal strength, battery levels, interference, and traffic patterns. These systems can self-heal around failed nodes, balance network traffic, and adjust to changing environmental conditions. When combined with energy-aware routing algorithms that consider battery levels and energy harvesting capabilities, mesh networks can significantly extend the operational lifetime of IoT systems in challenging environments.

Figure 3 illustrates the three key elements of Connectivity Orchestration for next-generation IoT systems: Multi-Rat Management: This section shows how IoT devices intelligently manage multiple Radio Access Technologies (RATs). The central blue box represents an "IoT Device" with a "multi-RAT Manager" that coordinates connections across four different wireless technologies: Cellular For wide-area mobile connectivity, Wi-Fi For local high-bandwidth connections; BLE (Bluetooth Low Energy) For short-range, energy-efficient connections and LoRaWAN: For long-range, low-power wide area networking (Armentano et al., 2017). Dynamic Spectrum Access**:** This depicts how IoT systems can efficiently identify and utilize available radio spectrum. The colored bands represent different frequency bands, with the "available spectrum" highlighted. Below this is "Software Defined Radio" technology that enables devices to dynamically access and use these spectrum resources, adapting to local conditions and regulations. Mesh Network Resilience: This illustrates a mesh network topology where devices connect to multiple neighboring devices rather than directly to a central hub. The diagram shows Multiple nodes (circles) interconnected through various paths, A connection failure marked with an "X," and Alternative routing paths that maintain network connectivity despite the failure.

Together, these three technologies create robust, adaptive connectivity for IoT deployments, enabling them to maintain reliable connections across diverse environments while optimizing power usage, bandwidth, and resilience.

Figure 3: *Connectivity Orchestration Framework for Next-Generation IoT Networks.*



The integration of these connectivity orchestration strategies—Multi-RAT management, dynamic spectrum access, and resilient mesh networking—creates IoT communication fabrics that are more adaptable and reliable than ever before. These systems maintain connectivity in challenging conditions, optimize power and bandwidth use, and automatically adapt to environmental changes. As IoT deployments continue to scale, these orchestration capabilities will be critical in maintaining performance and reliability and minimizing operational costs.

Conclusion

This chapter has explored the evolution of IoT ecosystems, from their early forms as simple sensor networks to today's sophisticated, intelligent systems that integrate seamlessly with the physical world. We have traced how core IoT architectures have shifted—from centralized, cloud-reliant designs to distributed intelligence models that utilize computing resources throughout the entire network.

At the heart of next-generation IoT are three transformative paradigms: the edge-to-cloud continuum, cognitive IoT systems, and digital twin integration. Edge-to-cloud frameworks support the flexible distribution of computational tasks, optimizing for latency, bandwidth, and energy use. Cognitive IoT systems embed AI throughout the stack, from sensor-level processing to edge analytics and cloud-based deep learning. Digital twins provide real-time, evolving digital replicas of physical assets (Rani et al., 2024) and systems, enabling simulation, optimization, and predictive maintenance.

These paradigms work in concert—not in isolation—to create IoT ecosystems capable of more than just sensing and analysis. They empower systems to anticipate, adapt, and act autonomously. This shift from reactive to proactive operation marks a defining trait of next-generation IoT.

Connectivity innovations are the enablers of this transformation. Emerging technologies such as ambient IoT, zero-power devices, and orchestrated connectivity allow IoT networks to scale from thousands to billions of endpoints. Techniques like energy harvesting, backscatter communication, and passive sensing support ultra-low-power or even self-sustaining devices. At the same time, multi-RAT management, dynamic spectrum access, and resilient mesh networks build robust communication infrastructures that thrive across diverse and dynamic environments.

This foundational perspective will provide critical context as we progress through the following chapters to examine real-world implementations and application domains. The shift from isolated connected devices to intelligent, responsive ecosystems is not merely a technological milestone—it reflects a fundamental reimagining of how digital and physical systems interact. This evolution continues to accelerate, propelled by breakthroughs in AI, communication, miniaturization, and sustainable technologies.

While significant challenges remain—in security, interoperability, data governance, and environmental impact—the foundational technologies outlined in this chapter offer the tools to overcome them.

References

Agarwal, P., Manekiya, M., Ahmad, T., Yadav, A., Kumar, A., Donelli, M., & Mishra, S. T. (2022). A survey on Citizens Broadband Radio Service (CBRS). *Electronics, 11*(23), 3985. https://doi.org/10.3390/electronics11233985

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials, 17*(4), 2347–2376. https://doi.org/10.1109/COMST.2015.2444095

Anand, A. (2024). The digital twin model-paving the way for a furustic mining sector. (Accees Date: 07.03.2025) Available on: https://www.bbrief.co.za/2024/03/04/the-digital-twin-model-paving-the-way-for-a-futuristic-mining-sector/?

Armentano, R., Bhadoria, R.S., Chatterjee, P., & Deka, G.C. (Eds.). (2017). The Internet of Things: Foundation for Smart Cities, eHealth, and Ubiquitous Computing (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315156026

Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. *Computer, 40*(12), 33-37. https://doi.org/10.1109/MC.2007.443

Berry, Randall and Hazlett, Thomas W. and Honig, Michael and Laneman, J. Nicholas, Evaluating the CBRS Experiment (August 1, 2023). Available at SSRN: https://ssrn.com/abstract=4528763 or http://dx.doi.org/10.2139/ssrn.4528763

Chen, Y., Li, Y., Gao, M. *et al.* Throughput optimization for backscatter-and-NOMA-enabled wireless powered cognitive radio network. *Telecommun Syst* 83, 135–146 (2023). https://doi.org/10.1007/s11235-023-01012-6

Demertzi, V., Demertzis, S., & Demertzis, K. (2023). An overview of cyber threats, attacks, and countermeasures on the primary domains of smart cities. *Applied Sciences, 13*(2), 790. https://doi.org/10.3390/app13020790

Khan, L. U., Saad, W., Niyato, D., Han Z., and Hong, C. S. "Digital-Twin-Enabled 6G: Vision, Architectural Trends, and Future Directions," in *IEEE Communications Magazine*, vol. 60, no. 1, pp. 74–80, January 2022, doi: 10.1109/MCOM.001.21143

Khan, L. U., Yaqoob, I., Tran, N. H., Han, Z., & Hong, C. S. (2020). Network slicing: Recent advances, taxonomy, requirements, and open research challenges. IEEE Access, 8, 36009–36028. https://doi.org/10.1109/ACCESS.2020.2975072

Lee, K., & Man, K. L. (2022). Edge computing for Internet of Things. *Electronics, 11*(8), 1239. https://doi.org/10.3390/electronics11081239

Liu, V., Parks, A., Talla, V., Gollakota, S., Wetherall, D., & Smith, J. R. (2013). Ambient backscatter: Wireless communication out of thin air. In Proceedings of the ACM SIGCOMM 2013 conference on (SIGCOM '13). Association for Computing Machinery, New York, NY, USA, 39–50. https://doi.org/10.1145/2486001.2486015

Merenda, M., Porcaro, C., & Iero, D. (2020). Edge machine learning for AI-enabled IoT devices: A review. *Sensors, 20*(9), 2533. https://doi.org/10.3390/s20092533

Moysiadis, V., Sarigiannidis, P., & Moscholios, I. (2018). Towards distributed data management in fog computing. Wireless Communications and Mobile Computing, 2018. https://doi.org/10.1155/2018/7597686

Rani, S., Bhambri, P., Kumar, S., Pareek, P.K., & Elngar, A.A. (Eds.). (2024). AI-Driven Digital Twin and Industry 4.0: A

Conceptual Framework with Applications (1st ed.). CRC Press. https://doi.org/10.1201/9781003395416

Ray, P. P. (2016). A survey on Internet of Things architectures. *Journal of King Saud University-Computer and Information Sciences, 30*(3), 291–319. https://doi.org/10.1016/j.jksuci.2016.10.003

Wang, J., Vasisht, D., & Katabi, D. (2014). RF-IDraw: Virtual touch screen in the air using RF signals. In Proceedings of the ACM SIGCOMM 2014 conference on (SIGCOM '14). Association for Computing Machinery, New York, NY, USA, 235-246. https://doi.org/10.1145/2619239.2626330.

Yang, Z., Chen, M., Wong, K. K., Poor, H. V., & Cui, S. (2022). Federated learning for 6G: Applications, challenges, and opportunities. *Engineering, 8*, 33-41. https://doi.org/10.1016/j.eng.2021.12.002

Sethi, P., & Sarangi, S. R. (2017). Internet of Things: Architectures, protocols, and applications. *Journal of Electrical and Computer Engineering, 2017*, 9324035. https://doi.org/10.1155/2017/9324035

CHAPTER 3

INTELLIGENCE, SECURITY, AND SUSTAINABILITY İN NEXT-GENERATION IOT

ÖZLEM BATUR DİNLER[1]

Introduction

The evolution of the Internet of Things (IoT), from simple connected devices to complex, intelligent ecosystems, has enabled transformative capabilities across industries. However, this rapid growth has also introduced a range of technical and societal challenges. As IoT systems expand to encompass billions of interconnected devices, three foundational elements emerge as critical for ensuring their effective and responsible deployment: distributed intelligence, robust security, and sustainable design.

Distributed intelligence enables IoT systems to move beyond reliance on centralized cloud platforms by incorporating decision-making capabilities across various layers of the network. This approach improves system responsiveness, reduces latency, enhances data privacy, and increases resilience. At the same time, security has become a major concern, especially as IoT technologies are integrated into sensitive domains such as healthcare, industrial

[1] Assistant Professor, Siirt University, Department of Computer Engineering, 0000-0002-2955-6761

systems, and critical infrastructure. The diversity of devices and the open nature of IoT networks expose these systems to vulnerabilities that traditional security methods cannot sufficiently address. New security models based on zero-trust principles and hardware-level protections are emerging as essential components in safeguarding these environments.

Finally, as IoT device numbers increase, their environmental impact expands accordingly. The energy consumption, resource usage, and e-waste associated with IoT must be addressed through sustainable design practices. By adopting energy-efficient architectures and promoting device reuse, repair, and recycling, IoT systems can become more environmentally and economically sustainable.

This chapter explores these three pillars—intelligence, security, and sustainability—as interconnected forces that will shape the future of IoT. Understanding and integrating these dimensions is essential not only for technological progress but also for meeting the broader societal and ecological demands of the future.

Edge AI and Distributed Computing for IoT

Federated Learning at the Edge

Traditional machine learning approaches in IoT environments primarily rely on centralized model training. Data collected from distributed devices is transmitted to cloud platforms for processing. While effective in some cases, this approach presents challenges such as bandwidth limitations, latency requirements, privacy regulations, and security concerns. Distributed devices can train models without centralizing raw data through federated learning, providing an attractive alternative approach.

The fundamental principle of federated learning is that model updates, rather than raw data, are shared (Reddy et al., 2025) between edge devices and coordination servers. Typically, an initial model is deployed to participating edge devices, which train the model using locally available data. Instead of transmitting sensitive raw data, devices send only model parameter updates to a central server (Li et al., 2024). This server aggregates updates from multiple devices to create an improved global model (Rani and Taneja, 2024), which is then redistributed to the edge. This iterative process allows the model to improve over time while ensuring data remains local (Alharbey and Jamil et al., 2025).

Advantages of Federated Learning in IoT

- **Privacy Enhancement:** Since raw data never leaves the originating devices, federated learning minimizes exposure to sensitive information. This is particularly valuable in applications such as healthcare monitoring, industrial process optimization, and consumer behavior analysis.
- **Bandwidth Efficiency:** Model updates are significantly smaller than raw datasets, reducing network congestion and improving efficiency.
- **Latency Reduction:** Locally trained models can make predictions without requiring round-trip communication to cloud platforms, enabling real-time applications even in environments with intermittent connectivity.

Challenges and Solutions in Federated Learning for IoT

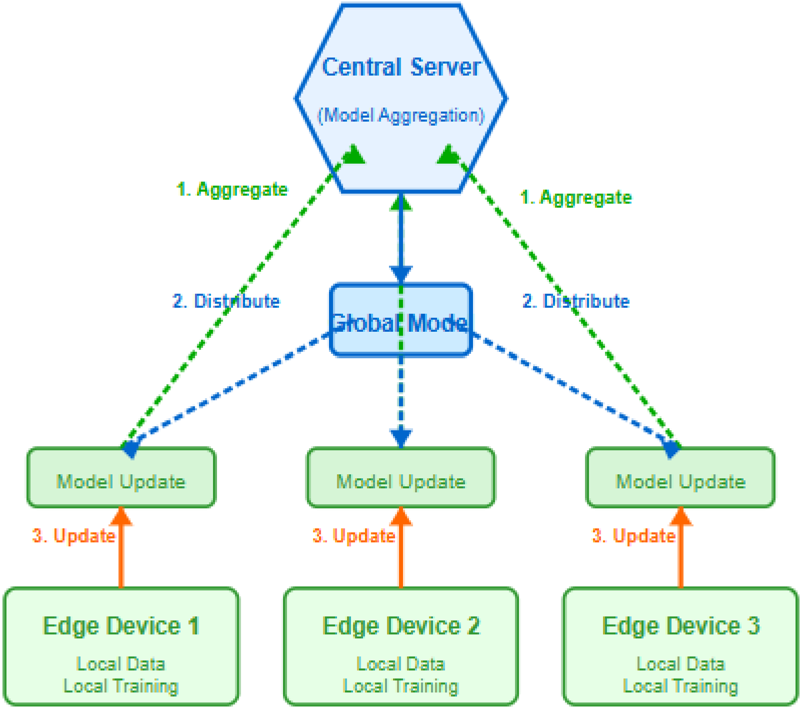
Implementing federated learning in IoT environments presents unique challenges due to the heterogeneity of edge devices (Revathi et al., 2024). Unlike cloud-based training environments with uniform computational resources, IoT devices vary significantly in processing capabilities, memory, and energy constraints. Specialized techniques are required to address these challenges:

- **Heterogeneous Device Optimization:** Techniques such as knowledge distillation (creating compact "student" models from larger "teacher" models), quantization (reducing model precision to decrease computational requirements), and pruning (eliminating non-essential network connections) enable devices with different capabilities to participate in federated learning.
- **Adaptive Model Partitioning:** By distributing neural network components across devices, edge nodes, and cloud platforms, computational efficiency is enhanced. For example, feature extraction layers can run on edge devices, while more complex transformations occur on fog nodes or cloud servers. This dynamic partitioning optimizes resource utilization based on network conditions, battery levels, and processing requirements.

Machine learning models can be trained across distributed devices without centralizing raw data using Federated Learning, as depicted in Figure 1. The workflow comprises three key processes: (1) Aggregation (Green Arrows) - Edge devices train models locally and send only model updates to the Central Server, which combines these updates to enhance the global model; (2) Distribution (Blue Arrows) - The improved Global Model is sent back to all participating edge devices; (3) Updating (Orange Arrows) - Edge devices 1, 2, and 3 use their local data to train the model and generate updates, with each device maintaining its own local data and conducting training independently.

This approach offers several advantages for IoT systems: it enhances privacy by keeping sensitive data on local devices; it reduces bandwidth usage by transmitting only model parameters instead of raw data; and enables devices to (Berkani et. al.,) make predictions locally without requiring constant connectivity to a central cloud platform. The architecture creates a learning system where intelligence is distributed across the network while maintaining data privacy and minimizing communication overhead.

Figure 1: *Federated learning.*



Integrating federated learning into IoT ecosystems enables a new generation of intelligent applications. Smart home systems can learn user preferences without exposing sensitive behavioral data. Industrial equipment can develop customized anomaly detection models based on local operating conditions while maintaining proprietary process security. Agricultural systems can optimize irrigation and fertilization strategies based on hyperlocal conditions while aggregating insights across multiple farms. By keeping data local while sharing insights, federated learning addresses many of the privacy, bandwidth, and latency challenges that have traditionally hindered AI adoption in distributed IoT environments.

Swarm Intelligence for Autonomous IoT

As IoT deployments scale to thousands or even millions of interconnected devices, traditional centralized control approaches become increasingly impractical. Communication overhead, single points of failure, and scalability limitations hinder performance. Swarm intelligence provides an alternative paradigm inspired by biological systems such as ant colonies, bird flocks, and bee swarms. In swarm-based IoT systems, collections of devices exhibit emergent intelligence and solve complex problems through local interactions without centralized control.

Key Principles of Swarm Intelligence in IoT

- **Decentralized Decision-Making:** Rather than relying on a central controller to gather information, make decisions, and distribute commands, swarm-based approaches enable individual devices to act autonomously. Devices respond to environmental stimuli and communicate with nearby nodes, leading to global behaviors emerging from local interactions. For example, a swarm of temperature sensors could collectively identify thermal anomalies without a single device having a complete environmental picture. Similarly, autonomous drones could coordinate delivery routes dynamically without a centralized dispatcher.

- **Collective Perception:** In traditional sensor networks, each device operates independently, transmitting data to central points for analysis. In contrast, swarm-based systems allow devices to share observations with neighbors, creating a distributed awareness that surpasses individual device capabilities. This approach enables robust detection of complex phenomena. For instance, a swarm of acoustic sensors could more accurately triangulate sound sources, or autonomous vehicles could collectively map road conditions by sharing real-time observations.
- **Self-Healing Networks:** One of the most valuable applications of swarm intelligence is the ability to form self-healing networks. Traditional IoT networks are vulnerable to device failures or communication disruptions, which can compromise overall system functionality. Swarm-based approaches enable networks to autonomously reconfigure in response to failures, maintaining operational continuity without centralized intervention. Devices can dynamically adjust their behavior, reroute communications, reassign tasks, or modify coverage patterns to compensate for failures. This enhances resilience, particularly in mission-critical applications or harsh environments where device failures are common.

Implementing Swarm Intelligence in IoT

Implementing swarm intelligence in IoT systems requires careful design of interaction rules, communication protocols, and local decision-making algorithms. These elements must balance:

- **Autonomy vs. Coordination:** Ensuring devices operate independently while contributing to collective goals.
- **Resilience vs. Efficiency:** Maintaining adaptability without excessive redundancy or wasted resources.
- **Local Optimization vs. Global Objectives:** Enabling devices to make decisions that benefit both individual performance and overall system functionality.

Emerging techniques such as reinforcement learning allow devices to refine behaviors based on experience, evolutionary algorithms optimize interaction rules over time and trust mechanisms enhance security in decentralized environments.

Swarm intelligence is particularly valuable in applications involving mobile or widely distributed devices, dynamic environments, and mission-critical operations. Autonomous vehicle fleets can optimize movement and task allocation without centralized control. Environmental monitoring systems can adjust sampling rates based on distributed observations. Manufacturing systems can dynamically reconfigure production pathways in response to equipment failures or changing priorities. By distributing intelligence across device populations rather than concentrating it in central controllers, swarm-based approaches unlock new levels of autonomy, adaptability, and resilience in IoT systems.

Zero-Trust and Hardware-Based Security in IoT

Zero-Trust IoT Security

The distributed and heterogeneous nature of IoT ecosystems presents unique security challenges that traditional perimeter-based approaches fail to address. Early IoT security models relied on network segmentation and gateway protection, assuming that devices within protected zones could be trusted. However, this assumption becomes increasingly problematic as IoT deployments grow in complexity—incorporating diverse device capabilities, multiple stakeholders, and intricate supply chains. Zero-trust security models provide a more resilient alternative by assuming potential compromise and requiring continuous verification for every access request, regardless of its source.

Continuous Authentication

Traditional authentication models typically authenticate devices once during connection establishment, granting them ongoing trust. In contrast, zero-trust IoT environments enforce continuous authentication, ensuring that devices continually validate their identity and authorization using multiple factors. These factors generally fall into three categories: (1) Inherent device characteristics – such as hardware identifiers or cryptographic keys embedded during manufacturing. (2) Secured credentials – including digital certificates, passwords, or encrypted shared secrets. (3) Behavioral signatures – encompassing operational patterns, communication protocols, and usage rhythms.

Advanced implementations leverage behavioral authentication, which continuously monitors device operations for anomalies indicative of compromise. For example, if a smart thermostat unexpectedly attempts to access financial systems, it would trigger protective measures—even if it presents valid credentials.

Multi-Factor Behavioral Authentication

Multi-factor behavioral authentication strengthens security by analyzing device behavior over time to establish baseline profiles. Deviations from these profiles trigger additional verification or protective actions, even when conventional credentials remain valid. This approach is particularly effective for detecting sophisticated attacks where adversaries gain control of legitimate devices but use them in abnormal ways. Behavioral indicators may include

Communication patterns, data access requests, power consumption profiles, and timing characteristics.

By combining multiple behavioral indicators, systems can detect subtle anomalies that might indicate a breach while minimizing false positives that could disrupt legitimate operations.

Micro-Segmentation

Micro-segmentation enhances security by dividing IoT environments into granular security zones with independent access controls; unlike traditional network segmentation, which groups similar devices into broad zones, micro-segmentation creates much smaller zones—potentially down to individual devices—with precisely tailored access permissions. This approach restricts lateral movement by compromised devices, containing potential breaches within narrowly defined boundaries. For example, instead of placing all manufacturing sensors in a single security zone, micro-segmentation would create distinct zones for each sensor type, ensuring precise control over data flows.

Verifiable Computing

Verifiable computing ensures that edge processing occurs correctly, even on potentially compromised devices. As intelligence shifts from the cloud to the edge, ensuring computational integrity becomes critical. Verifiable computing methods allow devices to generate cryptographic proof that their computations were performed correctly—without exposing underlying data or requiring fully trusted execution environments. These techniques are especially valuable in federated learning applications, where devices train models locally on sensitive data. By providing cryptographic proof of execution, edge devices can securely participate in distributed intelligence systems without requiring complete trust in their internal operations.

By integrating continuous authentication, micro-segmentation, and verifiable computing, security architectures can effectively address the inherent vulnerabilities of distributed IoT environments. These approaches align with the reality of modern IoT deployments, where devices from multiple vendors, each with varying security capabilities, must securely interoperate in complex environments.

Hardware-Based Security

While software-based security measures provide essential protection for IoT systems, they ultimately depend on the integrity of the underlying hardware. Sophisticated attacks can compromise device firmware or operating systems, bypassing software-based defenses. Hardware-based security mechanisms mitigate this risk by embedding protection directly into the silicon, establishing a foundation of trust upon which software security measures can be built. These hardware roots of trust are particularly valuable in IoT environments, where devices may operate in physically accessible locations with minimal monitoring.

Physically Unclonable Functions (PUFs)

Physically Unclonable Functions (PUFs) exploit inherent variations in semiconductor manufacturing processes to generate unique device fingerprints that cannot be duplicated or simulated. Even chips produced from the same design exhibit slight physical differences due to variations in materials and fabrication conditions. PUFs leverage these differences to create device-specific cryptographic keys without storing them in memory, making them highly resistant to extraction. This approach is especially beneficial for lightweight IoT devices with limited secure storage capabilities, as cryptographic material is derived dynamically rather than stored explicitly.

Trusted Execution Environments (TEEs)

Trusted Execution Environments (TEEs) provide isolated processing domains that protect sensitive operations even if the main operating system is compromised. These secure enclaves use hardware-enforced isolation to create protected memory regions and execution environments that remain secure even if privileged system software is compromised. TEEs allow devices to securely store cryptographic keys, process sensitive data, and execute critical security functions in isolation.

Prominent TEE implementations include ARM TrustZone, Intel Software Guard Extensions (SGX), and AMD Secure Encrypted Virtualization (SEV) (Hong et al., 2018). These technologies are particularly valuable for edge devices performing sensitive local analytics or managing cryptographic material, as they provide hardware-enforced isolation beyond what software protections can achieve.

Secure Elements

Secure Elements are dedicated security chips designed for cryptographic operations and credential storage. These tamper-resistant hardware components include specialized features to resist physical attacks, such as shields against side-channel analysis, temperature and voltage monitoring to detect tampering, and encrypted memory to protect stored secrets.

Secure Elements often incorporate dedicated cryptographic accelerators, enabling efficient implementation of strong cryptography even on resource-constrained devices. Secure Elements provides robust protection for device identities, encryption keys, and authentication credentials by isolating critical security functions in dedicated hardware with limited attack surfaces.

Integrating hardware security technologies such as PUFs, TEEs, and Secure Elements creates a layered defense that addresses the unique security challenges of IoT environments. PUFs establish device-specific identities based on physical characteristics, TEEs enable secure processing even in compromised software environments, and Secure Elements protect cryptographic material with dedicated hardware. Together, these technologies form a hardware foundation for security that complements software-based protections, enabling comprehensive security architectures for next-generation IoT deployments.

As IoT devices become increasingly embedded in critical infrastructure, industrial systems, and sensitive personal applications, hardware-based security will be crucial in ensuring resilience against sophisticated cyber threats. By incorporating these technologies from the design phase, manufacturers can create devices with enduring security foundations that remain robust throughout their operational lifetimes, even as new software vulnerabilities emerge.

Data Management for IoT Scale

Semantic Interoperability

The heterogeneous nature of IoT ecosystems presents significant challenges for data integration and interoperability (Rewathi et al., 2024). Different manufacturers' devices employ diverse protocols and data models, resulting in syntactic incompatibility of their information despite representing similar concepts. Traditional point-to-point integration approaches become increasingly unmanageable as ecosystems grow, limiting the ability to extract meaningful insights. Semantic interoperability addresses this challenge by enabling a shared understanding of data meaning across diverse systems.

Knowledge Graphs for IoT

Knowledge graphs provide a unified representation of devices, their capabilities, relationships, and generated data. These graph-based models capture not only data structures but also the semantic meaning of entities and their interconnections, facilitating intelligent queries and inference. For instance, a knowledge graph can represent a temperature sensor in a specific room, detailing its measurement unit, accuracy, and reporting frequency. By incorporating domain-specific ontologies—such as those for industrial automation, healthcare, or smart buildings—knowledge graphs enhance semantic understanding within specialized applications.

Ontology Alignment

Ontology alignment complements knowledge graphs by mapping different semantic models, enabling interoperability across vendors, domains, and applications. Instead of enforcing uniform data models—impractical in diverse ecosystems—ontology alignment bridges different representations through semantic equivalence. Machine learning techniques help identify mappings between concepts, properties, and relationships across ontologies, with varying levels of human verification. For example, an alignment system might determine that one vendor's "ambient_temp" property is semantically equivalent to another's "room_temperature," allowing seamless integration without manual mapping.

Self-Describing Data

Self-describing data formats extend semantic capabilities by embedding contextual metadata within the data itself. Instead of relying on external schemas, these formats use technologies such as JSON-LD (JavaScript Object Notation for Linked Data) and RDF (Resource Description Framework) to encode semantic annotations.

A temperature reading, for instance, might include links to definitions specifying the unit of measurement, sensor type, calibration data, and physical location. This approach enables independent and correct data interpretation across different systems, facilitating ad-hoc analysis and unforeseen applications.

Figure 2 illustrates the concept of Semantic Interoperability for IoT systems. This diagram shows how heterogeneous IoT devices from different vendors can communicate meaningfully through semantic technologies. The figure is organized into three main sections: (1) Shows diverse IoT devices from different vendors (A, B, C, and D) that initially use incompatible data formats and protocols. (2) Displays three key semantic technologies that enable interoperability:

- Knowledge Graphs: A graph structure showing how IoT concepts like "Room," "Sensor," "Data," "Unit," and "Time" are interconnected, creating a unified understanding of relationships between devices and their data.

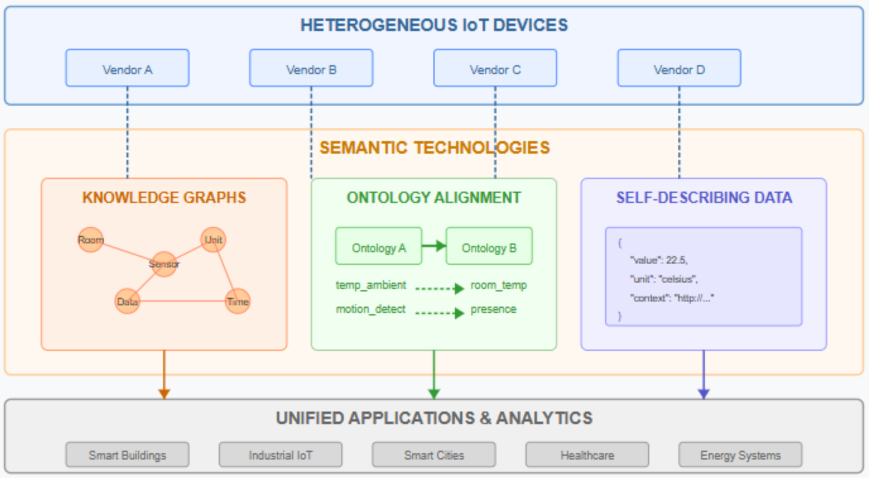
- Ontology Alignment: Demonstrates how terms from different vendors can be mapped to each other (e.g., "temp_ambient" = "room_temp" and "motion_detect" = "presence"), allowing systems to understand equivalent concepts.

- Self-Describing Data: This shows how data can carry its own semantic meaning through formats like JSON-LD, where each value includes context information about what it represents.

(3) Illustrates how these semantic technologies enable unified applications and analytics across different domains, including smart buildings, industrial IoT, smart cities, healthcare, and energy Systems.

This framework is crucial for IoT environments where devices from multiple manufacturers need to share and integrate data meaningfully. It enables more sophisticated applications that can work across traditionally siloed systems.

Figure 2: *Semantic interoperability framework for heterogeneous IoT ecosystems.*



```
{
    "value": 22.5,
    "unit": "celsius",
    "context": "http://..."
}
```

IoT ecosystems achieve greater flexibility and interoperability by integrating knowledge graphs, ontology alignment, and self-describing data. Applications can dynamically discover relevant data sources, interpret information correctly, and combine insights across traditionally siloed systems. This is particularly valuable in complex environments like smart cities, where diverse data streams—ranging from transportation to environmental monitoring—must be integrated for comprehensive urban management.

Time-Series Optimization

Time-series data—sequential sensor measurements with timestamps—constitutes the primary data type in IoT deployments. With thousands or even millions of sensors generating high-frequency data points, efficient data management becomes a critical challenge. Next-generation IoT systems employ specialized techniques to optimize storage, transmission, and analysis of time-series data.

Adaptive Sampling

Adaptive sampling dynamically adjusts measurement frequency based on signal characteristics, balancing data resolution with resource efficiency. Traditional fixed-rate sampling either collects excessive data during stable periods (wasting resources) or misses critical variations during rapid changes. Adaptive techniques mitigate this by:

- Increasing frequency during rapid fluctuations to capture essential transitions.
- Reducing frequency during stable periods to conserve bandwidth and storage.
- Utilizing compressive sensing, which reconstructs signals from sparse samples.
- Applying delta encoding, which records only changes rather than absolute values.
- Incorporating predictive sampling, which models expected behavior and logs deviations.

These approaches significantly reduce data volumes while preserving essential information and improving storage, transmission, and processing efficiency.

Temporal Compression

Temporal compression algorithms optimize the storage and transmission of time-series data by leveraging domain-specific patterns. Unlike generic compression methods, which yield modest gains, temporal compression exploits unique time-series characteristics such as:

- Piecewise approximation, which represents signal segments with mathematical functions.
- Symbolic aggregation, converting numerical series into symbolic representations.
- Pattern dictionaries, identifying and encoding recurring sequences efficiently.

By utilizing these methods, IoT systems achieve compression ratios exceeding 100:1 in many applications, particularly in environments with stable or cyclical parameters.

Distributed Time-Series Processing

As IoT data volumes grow, centralized time-series processing becomes impractical due to bandwidth, latency, and cost constraints. Distributed architectures address this by processing data across the edge-fog-cloud continuum: Edge devices handle local aggregation and filtering, fog nodes conduct intermediate analytics at regional levels, and cloud platforms manage global analysis and long-term storage.

Specialized time-series databases support high-speed ingestion, efficient compression, and fast analytical queries. Advanced systems execute distributed queries across processing tiers, routing workloads based on data locality and resource availability. This ensures real-time analytics while minimizing data movement and optimizing computational resources.

Sustainable IoT Design

As the Internet of Things (IoT) expands, its energy consumption and environmental impact become increasingly significant concerns. Billions of connected devices require power, and their manufacturing, operation, and disposal contribute to resource depletion and electronic waste. Sustainable IoT design minimizes energy consumption, extends device lifespans, and reduces environmental footprints. This can be achieved through energy-aware architectures and circular economy principles, ensuring IoT systems remain efficient and environmentally responsible.

Energy-Aware Architecture

Managing power consumption is critical for large-scale IoT deployments, especially in battery-powered and remote devices. Traditional architectures often consume energy inefficiently, leading to unnecessary waste. Next-generation IoT systems incorporate energy-aware designs that optimize power usage across hardware and software layers, improving performance and sustainability.

Power-Proportional Computing

Power-proportional computing ensures that energy consumption scales with computational load, reducing power usage during idle or low-activity periods. Traditional computing systems often consume significant power, even when underutilized, due to inefficient power states and always-on components.

Energy-efficient designs address this through:

- Dynamic voltage and frequency scaling (DVFS): Adjusting processor performance based on workload demand.
- Aggressive sleep states: Powering down unused components when not needed.
- Workload consolidation: Grouping tasks to maximize efficiency and enable deeper sleep states when idle.

Workload Scheduling for Energy Harvesting

In energy-harvesting IoT systems, devices rely on renewable sources such as solar, thermal, or kinetic energy. Since power availability fluctuates, these systems require intelligent scheduling to align computational tasks with energy availability. Techniques include:

- **Energy forecasting:** Predicting future energy availability based on historical data and environmental conditions.
- **Progressive precision computing:** Adjusting computational accuracy based on available energy.
- **Opportunistic computing:** Executing complex operations incrementally during energy-rich periods.

By optimizing workload scheduling, energy-harvesting IoT systems can operate autonomously in remote areas where conventional power sources are impractical, such as environmental monitoring stations and agricultural sensors.

Thermal-Aware Design

As processing power increases, so does heat generation, which can lead to performance degradation and higher energy consumption due to cooling requirements. Traditional thermal management techniques often rely on reactive throttling, which reduces system efficiency. Advanced thermal-aware designs incorporate:

- **Heterogeneous multiprocessing:** Distributing workloads across energy-efficient and high-performance cores based on thermal conditions.
- **Task migration:** Shifting computations to cooler parts of the system to prevent localized overheating.

- **Proactive thermal management:** Using predictive models to optimize workload distribution before overheating occurs.

For edge computing infrastructure, improved thermal management reduces cooling demands, which typically account for 30-50% of total data center energy consumption. IoT deployments achieve greater efficiency by integrating these strategies while maintaining long-term reliability.

Circular IoT Economy

Beyond energy consumption, IoT sustainability also depends on how devices are manufactured, maintained, and disposed of. Traditional electronics follow a linear model—devices are produced, used, and discarded, generating substantial electronic waste. Product design focused on repair, upgradeability, and responsible disposal creates a circular economy that reduces waste while optimizing resource efficiency.

Modular Device Architecture

Modular design enables selective component replacement, repair, and upgrades, extending device lifespans and reducing waste. Instead of discarding entire devices when one component fails, modular IoT devices allow for targeted repairs. Key features of modular IoT design include:

- **Hot-swappable sensor modules:** Sensors can be replaced without disrupting overall device operation.
- **Stackable processing and connectivity layers:** Devices can be upgraded (e.g., from 5G to 6G) without replacing the entire system.

- **Standardized power and data interfaces:** Components remain compatible across multiple product generations.

By shifting from monolithic to modular designs, IoT manufacturers can reduce material waste while allowing users to extend device functionality over time.

Biodegradable Electronics

Traditional electronics contain materials that persist in the environment for decades, contributing to pollution and toxic waste. Biodegradable electronics offer a sustainable alternative by using materials that naturally decompose under controlled conditions. Innovations in biodegradable electronics include:

- **Cellulose-based substrates:** Derived from plant materials, these decompose naturally after disposal.
- **Carbon-based conductive inks:** Avoid heavy metals typically found in circuit boards.
- **Bioplastic and mycelium-based enclosures:** Providing sustainable casings for temporary or disposable IoT devices.

While current biodegradable electronics may not match the performance of traditional components, they are increasingly viable for short-term applications, such as environmental monitoring or agricultural sensors.

Digital Passports for IoT Devices

One of the biggest challenges in recycling IoT devices is the lack of visibility into their internal components. Digital passports address this issue by maintaining a secure, detailed record of a device’s materials, usage history, and end-of-life handling instructions. These records typically include:

- **Bill of materials (BoM):** Listing all components and their origins.
- **Repair and upgrade history:** Tracking modifications made during a device's lifetime.
- **Environmental compliance data:** Ensuring adherence to sustainability regulations.

Digital passports create transparent and tamper-proof records by leveraging blockchain or distributed ledger technologies. This approach benefits manufacturers, recyclers, and regulatory bodies by enhancing supply chain accountability and facilitating more effective recycling and refurbishment efforts.

As IoT adoption accelerates, sustainability must be a priority. Energy-aware architectures, including power-proportional computing, energy-harvesting workload scheduling, and thermal-aware designs, help minimize energy consumption and improve device efficiency. At the same time, circular economy principles, such as modular design, biodegradable electronics, and digital passports, reduce electronic waste and extend device lifespans.

By integrating these sustainability strategies, IoT systems can meet growing performance demands and reduce their environmental impact, ensuring a more responsible and efficient digital future.

Real-World Applications of Next-Generation IoT

Precision Agriculture

Smart farming deploys IoT technology to enhance agricultural efficiency, boost harvest production, and reduce ecological footprint.

- **Environmental Intelligence**: Arrays of energy-neutral sensors provide constant surveillance of soil parameters, climate conditions, and plant vitality, delivering immediate data for decision-making.

- **Autonomous Operations**: Edge AI-enabled drones, equipped with spectral analysis capabilities, conduct aerial surveys, complementing ground-level sensor data. Federated learning systems optimize irrigation, fertilization, and other farming operations based on field data.
- **Digital Twin Integration**: Digital twins synthesize data from various sources to simulate different scenarios and predict outcomes, enabling farmers to make informed decisions.
- **Sustainability and Carbon Monitoring**: IoT systems enable carbon sequestration monitoring and reporting, allowing farmers to participate in carbon markets and generate additional revenue through sustainable practices.

The integration of these technologies creates a sophisticated precision agriculture ecosystem that optimizes production, enhances environmental stewardship, and promotes economic sustainability.

Autonomous Industrial Operations

Next-generation IoT is driving a paradigm shift in manufacturing, enabling increasingly autonomous operations with enhanced efficiency, safety, and adaptability.

- **Swarm Robotics:** Swarm-based systems with autonomous mobile robots (AMRs) and collaborative robots (cobots) enable dynamic task allocation and production line reconfiguration without centralized control.

- **Digital Thread Integration:** Digital thread implementations create continuous information flows between design, production, and operation phases, enabling real-time root cause analysis, proactive intervention, and continuous improvement cycles.
- **Zero-Trust Security:** Zero-trust security architectures with hardware roots of trust ensure robust protection for increasingly connected and automated industrial systems.
- **Semantic Interoperability:** Standardized information models and ontology-based semantic representations facilitate seamless integration of mixed-vendor equipment, reducing integration costs and enabling flexible reconfiguration.
- **Predictive Maintenance:** Energy harvesting sensors and edge analytics enable comprehensive machine health monitoring and predictive maintenance, minimizing downtime and optimizing operational efficiency.

The convergence of these technologies creates intelligent and resilient manufacturing ecosystems capable of autonomous operation, continuous optimization, and adaptation to changing conditions.

Smart Cities

IoT plays a crucial role in developing smart cities, enabling more efficient urban management, improved services, and enhanced quality of life for citizens.

- **Smart Traffic Systems**: IoT sensors combined with AI technologies streamline traffic patterns, minimize congestion, and boost transportation performance.

- **Intelligent Energy Networks**: IoT facilitates instantaneous tracking and control of energy usage, enhancing distribution efficiency and supporting renewable energy integration.
- **Environmental Surveillance**: IoT devices track air and water conditions, sound pollution, and other ecological indicators, generating insights for strategic planning and timely interventions.
- **Community Protection**: IoT technology strengthens public security through monitoring systems, emergency response coordination, and anticipatory law enforcement measures.

Healthcare

The IoT is significantly transforming the healthcare sector by enabling continuous remote patient monitoring, facilitating the implementation of personalized medicine, and enhancing the overall efficiency and responsiveness of healthcare delivery systems.

- **Remote Patient Monitoring**: Wearable sensors and connected devices allow continuous monitoring of patient's vital signs and health conditions (Sharma et al., 2024)., enabling timely interventions and reducing hospital readmissions.
- **Smart Hospitals**: IoT technologies optimize hospital operations, improve patient flow, and enhance the efficiency of medical equipment management.
- **Personalized Medicine**: IoT devices collect and analyze patient data to tailor treatments and therapies to individual needs (Sharma et al., 2024) , improving outcomes and reducing side effects.

These are just a few examples of how next-generation IoT transforms various sectors. As these technologies continue to advance, it is anticipated that increasingly innovative applications will emerge, aiming to address critical challenges in various domains and significantly enhance the quality of human life.

Future Directions and Research Challenges

Three transformative frontiers are shaping the next generation of IoT ecosystems: quantum IoT integration, biological and molecular IoT interfaces, and neuromorphic computing. These emerging technologies promise to enhance security, efficiency, and adaptability across IoT applications, paving the way for a new era of intelligent and interconnected systems (Rani and Taneja, 2024). However, significant research and development challenges must be addressed before these innovations reach their full potential.

Quantum IoT Integration

Integrating quantum computing with IoT holds immense potential, particularly in three key areas: quantum-resistant cryptography, quantum sensing, and quantum computing for optimization. As quantum computing advances, it poses a threat to traditional cryptographic methods used to secure IoT devices and networks. This necessitates the development of quantum-resistant cryptographic algorithms capable of withstanding attacks from quantum adversaries, ensuring long-term data security in IoT applications (Garg et al., 2024).

Beyond security, quantum sensing represents a groundbreaking innovation for IoT. These sensors offer unparalleled measurement precision, which could revolutionize fields such as medical diagnostics, environmental monitoring, and industrial automation. These devices can detect minute changes in temperature, pressure, and biological markers by leveraging quantum effects, significantly enhancing IoT's sensing capabilities.

Additionally, quantum computing can solve complex optimization problems in IoT ecosystems. Tasks such as resource allocation, real-time traffic management, and route optimization could be executed more efficiently through hybrid classical-quantum computing systems. Although still in its early stages, this integration promises to unlock new levels of computational power and efficiency. However, overcoming hardware limitations, error correction challenges, and scalability concerns remains crucial for the widespread adoption of quantum-IoT integration.

Biological and Molecular IoT Interfaces

The convergence of IoT with biological systems is opening new frontiers in healthcare, biotechnology, and environmental monitoring. Biocompatible sensors enable continuous, real-time health monitoring and provide unprecedented insights into physiological processes. Advanced systems are even capable of closed-loop interventions, where data collected by sensors can trigger automated therapeutic responses, offering new possibilities for personalized medicine and chronic disease management.

Beyond traditional electronics, molecular communication networks represent an innovative approach to IoT. These systems exchange information using chemical signals rather than electromagnetic waves, making them ideal for applications within living tissues. One of the most promising areas is targeted drug delivery, where molecular communication could enable precise and controlled release of medication at the cellular level, minimizing side effects and enhancing treatment effectiveness.

Biohybrid systems, which integrate living cells with electronic components, further expand the scope of IoT applications. These hybrid platforms could be used for environmental monitoring by detecting pollutants in water and soil or in medical diagnostics by interfacing directly with biological tissues. The ability to bridge the gap between electronic and biological systems represents a paradigm shift in how IoT interacts with the natural world. However, biocompatibility, stability, and scalability challenges must be addressed for widespread deployment.

Neuromorphic Computing for IoT

Neuromorphic computing, which draws inspiration from the structure and functionality of the human brain (Rani et al., 2024), presents a promising paradigm for enhancing the efficiency, adaptability, and intelligence of IoT systems (Singh et al., 2024). Traditional computing architectures struggle with the power and data constraints of edge devices, but neuromorphic systems address these limitations by mimicking biological neurons and synapses.

Spiking neural networks (SNNs) and neuromorphic hardware bring significant advantages to IoT by enabling real-time, low-power processing. Unlike conventional artificial neural networks, which rely on continuous data streams, SNNs operate on an event-driven basis, processing information only when changes occur. This reduces computational load and power consumption, making it ideal for resource-constrained IoT environments.

Event-based sensing enhances efficiency by generating data only when relevant changes are detected, reducing unnecessary transmission and storage. In addition, adaptive learning mechanisms enable IoT devices to constantly evolve and refine their performance by processing real-time data. This built-in learning ability ensures that the devices stay attuned to fluctuating conditions, enhancing both automation and decision-making in ever-changing environments.

Integrating neuromorphic computing with IoT represents a major step toward self-learning and autonomous edge intelligence. As research progresses, these advancements could lead to highly efficient, low-power IoT systems capable of real-time decision-making without reliance on cloud-based processing. However, challenges remain in hardware development, algorithm optimization, and large-scale implementation.

Breakthrough advancements in quantum computing, biological interfaces, and neuromorphic processing are shaping the future of IoT. These technologies have the potential to redefine the capabilities of IoT systems, making them more secure, precise, and adaptive. However, significant research challenges must be addressed to realize these innovations fully. As these fields evolve, interdisciplinary collaboration between engineers, biologists, and computer scientists will be essential in overcoming technological barriers and unlocking the full potential of next-generation IoT.

Conclusion

This chapter has examined the three foundational pillars shaping the future of next-generation IoT ecosystems: distributed intelligence, robust security, and sustainable design. These dimensions are not isolated; they are deeply interconnected, with progress in one area often reinforcing and necessitating advances in the others.

The shift toward distributed intelligence—enabled by federated learning, swarm intelligence, and edge AI—marks a critical evolution from traditional, cloud-centric IoT architectures. By moving computation closer to where data is generated, these systems achieve lower latency, greater privacy, and enhanced resilience. They also reduce the burden of data transmission and can continue functioning autonomously even in conditions with limited connectivity. As these approaches mature, we can expect increasingly complex behaviors to emerge from simple local rules, much like patterns found in nature.

At the same time, ensuring security in these increasingly autonomous systems is essential. Zero-trust security models, combined with hardware-based protections like PUFs, TEEs, and secure elements, form the backbone of next-generation IoT trust. These mechanisms support continuous verification and resilience against advanced threats—capabilities that are especially critical in high-stakes domains such as healthcare, industrial control, and infrastructure. As IoT devices gain autonomy, their ability to act safely and securely must be designed from the ground up.

Sustainability is becoming a non-negotiable requirement as IoT scales globally. With billions of devices deployed, energy-efficient design and responsible lifecycle management are vital. Techniques such as power-aware computing, thermal optimization, modular hardware design, and device reuse are key strategies for minimizing environmental impact. Circular economy principles—including repairability, recyclability, and eco-friendly materials—ensure IoT technologies can scale without unsustainable resource consumption or excessive e-waste.

Examples from agriculture, manufacturing, urban infrastructure, and healthcare illustrate how these three pillars combine to enable real-world, transformative applications. These are not just incremental improvements—they represent a new paradigm in how technology interacts with the physical world.

Emerging innovations such as quantum-enhanced IoT, brain-computer interfaces, and neuromorphic computing hint at even greater capabilities in sensing, intelligence, and energy efficiency. Though challenges remain, the trajectory is clear: IoT systems are evolving into increasingly intelligent, secure, and sustainable infrastructures.

Realizing this vision requires collaboration across technical, regulatory, ethical, and business domains. We can ensure that IoT technologies serve human and environmental well-being by embedding intelligence, security, and sustainability into the design process from the outset. This convergence marks not only the technical future of IoT but also its responsible and impactful path forward.

References

Alharbey, R. A., & Jamil, F. (2025). Federated learning framework for real-time activity and context monitoring using edge devices. *Sensors, 25*(4), 1266. https://doi.org/10.3390/s25041266

Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. *Computer, 40*(12), 33-37. https://doi.org/10.1109/MC.2007.443

Berkani MRA, Chouchane A, Himeur Y, Ouamane A, Miniaoui S, Atalla S, Mansoor W, Al-Ahmad H. Advances in Federated Learning: Applications and Challenges in Smart Building Environments and Beyond. *Computers*. 2025; 14(4):124. https://doi.org/10.3390/computers14040124

Dalenogare, L. S., Benitez, G. B., Ayala, N. F., & Frank, A. G. (2018). The expected contribution of Industry 4.0 technologies for industrial performance. *International Journal of Production Economics, 204*, 383-394. https://doi.org/10.1016/j.ijpe.2018.08.019

Demertzi, V., Demertzis, S., & Demertzis, K. (2023). An overview of cyber threats, attacks, and countermeasures on the primary domains of smart cities. *Applied Sciences, 13*(2), 790. https://doi.org/10.3390/app13020790

Dritsas, E., & Trigka, M. (2025). A survey on cybersecurity in IoT. *Future Internet, 17*(1), 30. https://doi.org/10.3390/fi17010030

Farahani, B., Barzegari, M., Aliee, F. S., & Shaik, K. A. (2020). Towards collaborative intelligent IoT eHealth: From device to fog, and cloud. *Microprocessors and Microsystems, 72*, 102938. https://doi.org/10.1016/j.micpro.2019.102938

Garg, V.D.G., Garg, J., Parasad, K.D., & Suneetha S. (2024). Future Outlook: Synergies Between Advanced AI and Cryptographic

Research. In Ruth, J., Vijayalakshmi, G., Visalakshi P., Uma, R. & A. Meenakshi (Eds.), *Innovative Machine Learning Applications for Cryptography* (pp. 27-46). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-1642-9.ch002

Hong, B., Kim, H.-Y., Kim, M., Suh, T., Xu, L., & Shi, W. (2017). FASTEN: An FPGA-based secure system for big data processing. *IEEE Design & Test, 35*(1), 30-38. https://doi.org/10.1109/MDAT.2017.2741464

Khan, L. U., Saad, W., Niyato, D., Han Z., and Hong, C. S. "Digital-Twin-Enabled 6G: Vision, Architectural Trends, and Future Directions," in *IEEE Communications Magazine*, vol. 60, no. 1, pp. 74–80, January 2022, doi: 10.1109/MCOM.001.21143

Khan, W. Z., Rehman, M. H., Zangoti, H. M., Afzal, M. K., Armi, N., & Salah, K. (2020). Industrial Internet of Things: Recent advances, enabling technologies and open challenges. *Computers & Electrical Engineering, 81*, 106522. https://doi.org/10.1016/j.compeleceng.2019.106522

Lee, K., & Man, K. L. (2022). Edge computing for Internet of Things. *Electronics, 11*(8), 1239. https://doi.org/10.3390/electronics11081239

Li, Y. *et al.* (2024). Federated Learning for Internet of Things. In: Donta, P.K., Hazra, A., Lovén, L. (eds) Learning Techniques for the Internet of Things. Springer, Cham. https://doi.org/10.1007/978-3-031-50514-0_3

Market Research. *Wearable sensors market analysis and forecast to 2031: By product type (temperature sensor, motion sensor, medical sensor, image sensor, position sensor, and others), application (eye wear, wrist wear, body wear, footwear, and others), and region*. Retrieved August 5, 2022, from: https://www.marketresearch.com/Global-Insight-Services-v4248/Wearable-Sensors-Forecast-Product-Type-32585677/

Merenda, M., Porcaro, C., & Iero, D. (2020). Edge machine learning for AI-enabled IoT devices: A review. *Sensors, 20*(9), 2533. https://doi.org/10.3390/s20092533

Moysiadis, V., Sarigiannidis, P., & Moscholios, I. (2018). Towards distributed data management in fog computing. Wireless Communications and Mobile Computing, 2018. https://doi.org/10.1155/2018/7597686

Perera, C., Liu, C. H., Jayawardena, S., & Chen, M. (2014). A survey on Internet of Things from industrial market perspective. *IEEE Access, 2*, 1660–1679. https://doi.org/10.1109/ACCESS.2015.2389854

Ramírez-Gordillo, T., Maciá-Lillo, A., Pujol, F. A., García-D'Urso, N., Azorín-López, J., & Mora, H. (2025). Decentralized identity management for Internet of Things (IoT) devices using IOTA blockchain technology. *Future Internet, 17*(1), 49. https://doi.org/10.3390/fi17010049

Rani, S., & Taneja, A. (Eds.). (2024). WSN and IoT: An Integrated Approach for Smart Applications (1st ed.). CRC Press. https://doi.org/10.1201/9781003437079

Reddy C, K.K., & Nag, A. (Eds.). (2025). Federated Learning for Neural Disorders in Healthcare 6.0 (1st ed.). CRC Press. https://doi.org/10.1201/9781003591085

Revathi, B., Hamsa, S., Shaik, N., Satpathy, S. K., Hari, & Myilsamy, S. (2024). Development of artificial intelligence of things and cloud computing environments through semantic web control models. In A. Ahmed Nacer & M. Abdmeziem (Eds.), *Emerging technologies for securing the cloud and IoT* (pp. 112–143). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0766-3.ch005

Sharma, L., & Garg, P.K. (Eds.). (2024). Deep Learning in Internet of Things for Next Generation Healthcare (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003451846

Singh, S., Sharma, P. K., Yoon, B., Shojafar, M., Cho, G. H., & Ra, I. H. (2020). Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city. *Sustainable Cities and Society, 63*, 102364. https://doi.org/10.1016/j.scs.2020.102364

Singh, J., Goyal, S., Kumar Kaushal, R., Kumar, N., & Singh Sehra, S. (Eds.). (2024). Applied Data Science and Smart Systems (1st ed.). CRC Press. https://doi.org/10.1201/9781003471059.

Tang, F., Kawamoto, Y., Kato, N., & Liu, J. (2020). Future intelligent and secure vehicular network toward 6G: Machine-learning approaches. *Proceedings of the IEEE, 108*(2), 292-307. https://doi.org/10.1109/JPROC.2019.2954595

Yang, Z., Chen, M., Wong, K. K., Poor, H. V., & Cui, S. (2022). Federated learning for 6G: Applications, challenges, and opportunities. *Engineering, 8*, 33-41. https://doi.org/10.1016/j.eng.2021.12.002

Zhao, Y., Zhao, J., Jiang, L., Tan, R., Niyato, D., Li, Z., Lyu, L., & Liu, Y. (2021). Privacy-preserving blockchain-based federated learning for IoT devices. In *IEEE Internet of Things Journal, 8*(3), 1817-1829. https://doi.org/10.1109/JIOT.2020.3017377

CHAPTER 4

MULTİLAYER SEGMENTATİON STRATEGY FOR MEANİNGFUL REGİON EXTRACTİON İN COMPLEX STRUCTURES

CANAN TASTIMUR[1]

Introduction

Image processing systems have become indispensable technological components, especially in the fields of computer vision and AI-supported analysis applications. Segmentation, one of the fundamental pillars of these systems, enables the separation of regions or objects of interest from the background within an image, thereby laying the groundwork for subsequent processing steps (Smith, 2018). The success of segmentation plays a critical role in visual analysis systems because it directly affects the accuracy of operations such as classification, defect detection, object recognition, and tracking (Johnson and Lee, 2019). Therefore, segmentation is not merely an image analysis technique but the cornerstone of extracting meaningful information from images.

The accuracy of segmentation often depends on the successful extraction of directional and structural details present in

[1] Assistant Professor, Erzincan Binali Yıldırım University, Department of Computer Engineering, Orcid: 0000-0002-3714-6826

the image. Many industrial, medical, and security-related images are characterized by complex patterns, multidirectional textures, and low-contrast surfaces (Kim and Park, 2020). In this context, traditional segmentation methods frequently experience performance degradation on such challenging data. Factors like noise, low resolution, irregular illumination, and texture-based variability can prevent segmentation algorithms from correctly distinguishing fine details (Chen et al., 2021). This leads to erroneous results, especially in data containing micro-level defects.

The clearest examples of this challenge appear in applications requiring high precision, such as medical imaging, textile industry, railway system safety, and metal surface analysis. For instance, inaccurate segmentation of tumor boundaries in a lung CT scan can directly impact clinical decisions (Wang and Zhang, 2017), while undetected defects in textile production can negatively affect product quality and customer satisfaction (Liu et al., 2019). Similarly, failure to timely and accurately detect cracks or wear in railway components can lead to serious safety issues (Kumar and Singh, 2018). In these application domains, segmentation is not just a technical procedure but is directly related to critical outcomes such as cost, safety, and human health.

Among frequently used segmentation techniques in the literature are thresholding, region growing, edge-based methods, clustering-based algorithms (e.g., FCM, K-Means), and convolutional neural network (CNN)-based deep learning approaches (Garcia and Lopez, 2020). However, many of these methods rely on predefined assumptions and do not always sufficiently represent the complex textures and patterns within images. Particularly, the lack of structural detail adversely affects the decision mechanisms of these algorithms, reducing segmentation quality (Patel and Shah, 2019). At this point, directional multi-scale preprocessing performed before segmentation provides significant advantages.

Accordingly, recent studies highlight multi-scale and directional preprocessing techniques such as the Redundant Contourlet Transform (RCT) (Do and Vetterli, 2005). These methods enhance horizontal, vertical, and diagonal edge information in the image, enabling segmentation algorithms to more accurately separate structural details (Chen et al., 2013). Preprocessing approaches like RCT both emphasize edges and combine information across different resolution levels, making small-scale anomalies more prominent (Zhang and Li, 2018).

The direct use of masks obtained after segmentation is often insufficient. The meaningful region extraction step following segmentation is another critical layer that determines system success. In this process, structures such as the Region Proposal Network (RPN), which have gained popularity in recent years, autonomously suggest potential regions of interest within the segmented image, preparing data for subsequent tasks such as classification and defect detection (Ren et al., 2015). Thus, it becomes possible to model not only pixels but also the potential objects or defects those pixels represent in a meaningful way.

In conclusion, the success of segmentation systems depends not only on the applied algorithm but also on the preprocessing techniques used before segmentation and the region proposal approaches employed afterward. The framework proposed in this study consists of multi-scale directional preprocessing similar to RCT, followed by Lattice Segmentation, and finally, RPN-based meaningful region proposal steps. This integrated approach aims to improve segmentation quality in applications such as surface defect detection, textile fault analysis, and microstructural deterioration identification in railway systems..

Directional and Multi-Scale Approach for Textile Defect Segmentatio

Image processing–based textile defect detection is a challenging problem, especially for automatically identifying very small, irregular, or low-contrast defects. In this study, a multi-stage preprocessing-segmentation approach is proposed to improve segmentation quality and to more effectively extract meaningful candidate regions. In the proposed method, first, the image details are enhanced from different directions using the directional multi-scale structure of RCT. Then, Lattice Segmentation is applied on these detailed images to accurately extract regional structures. In the final step, the segmentation results are supported by RPN architecture to automatically locate candidate defect regions. Thus, a more robust and automatable segmentation-detection system has been developed for small and difficult-to-detect textile defects where classical methods often fall short.

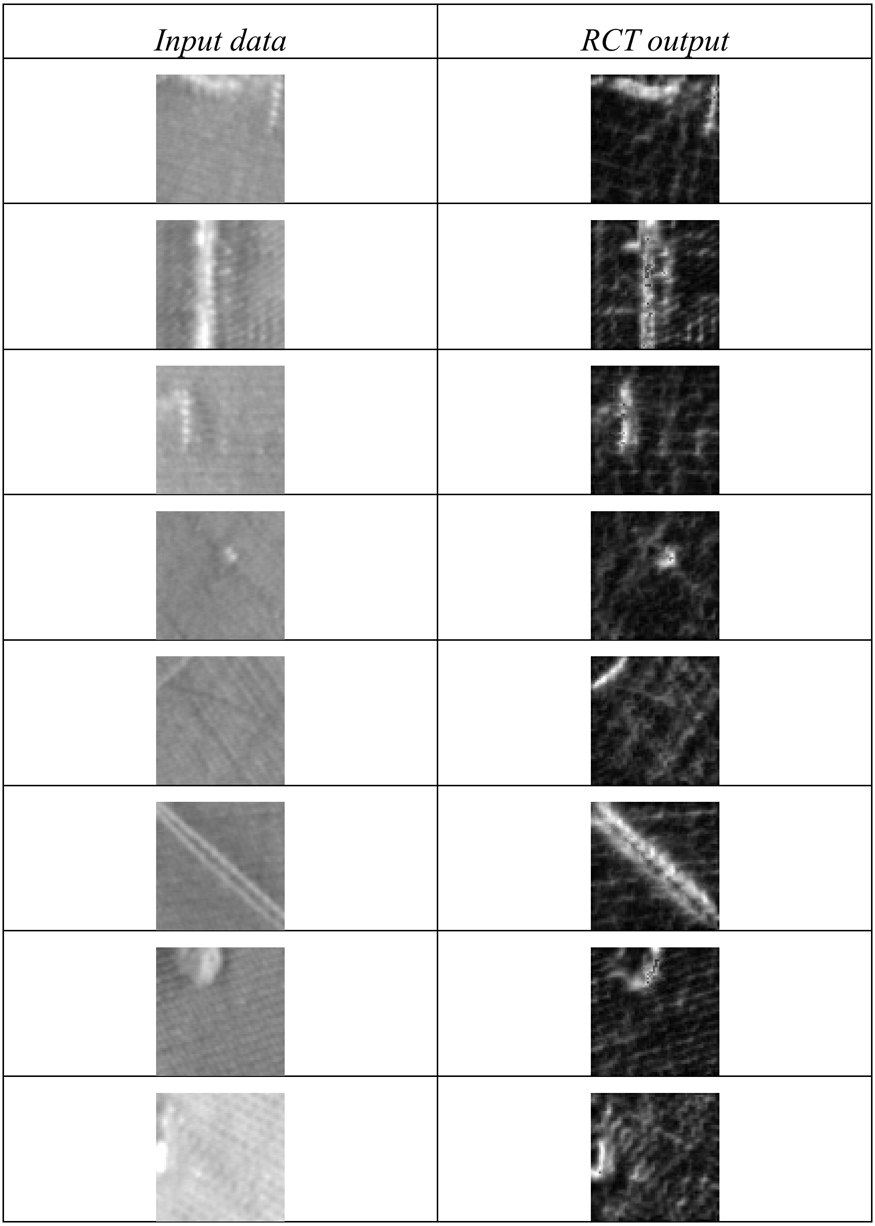
Redundant Contourlet Transform Directional Multi-Scale Preprocessing: In image processing and analysis, effectively extracting structural information such as edges and textures is of great importance. In this context, the Redundant Contourlet Transform (RCT) stands out as an effective method for multi-scale and multi-directional feature extraction. Unlike the classical wavelet transform, the contourlet transform provides a more natural and efficient representation of curved and directional structures in images (Do and Vetterli, 2005). However, the classical contourlet transform may cause some information loss due to its subsampling step. RCT eliminates this subsampling, providing a fully redundant transform that preserves detailed structural information (Chen et al., 2013).

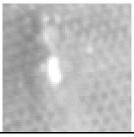
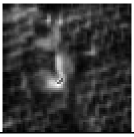
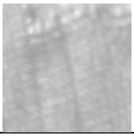
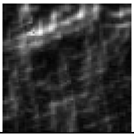
RCT-based directional multi-scale preprocessing aims to enhance structural features by filtering an image at multiple scales and directions. This method separately analyzes edge and texture components along horizontal, vertical, and diagonal directions. As a result, it achieves much richer and more detailed directional information compared to classical single-direction filtering methods (Zhang and Li, 2018). For example, fine lines, folds, or small defects on textile surfaces become more prominent through directional filtering. This directional filtering is a critical step in improving segmentation accuracy, especially for surfaces with complex textures (Wang and Chen, 2019).

Moreover, the multi-scale approach processes the image at various resolution levels. Typically, with pyramid-based multi-scale methods, the image is analyzed first at its original resolution, then at progressively reduced resolutions. Directional information extracted from each scale is then rescaled back to the original size and combined (Mallat, 2009). This process simultaneously captures large structural details at the macro level and fine structural elements at the micro level. Consequently, RCT multi-scale directional preprocessing enriches the image before segmentation, enabling the creation of more accurate and reliable masks.

This method also contributes significantly to industrial applications. For instance, in critical tasks such as detecting small yarn defects or surface irregularities in textile products, RCT highlights crucial image details and provides robust data for subsequent segmentation stages (Liu et al., 2020). Additionally, this technique is preferred in complex structural analyses in medical imaging, such as tumor or anomaly detection (Sun and Tang, 2021).

Figure 1 Application of RCT on Textile Data



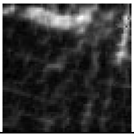
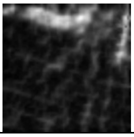
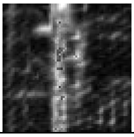
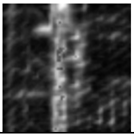
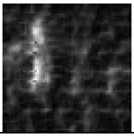
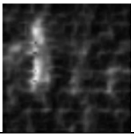
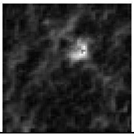
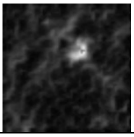
Effective Extraction of Directional and Structural Features with Lattice Segmentation: Lattice segmentation is an advanced method that aims to improve segmentation accuracy by modeling the structural organization and directional relationships within an image. It is particularly important to effectively utilize the spatial relationships between pixels in images with high textural and structural complexity (Zhao and Wang, 2017). Lattice-based approaches treat the image within a regular grid (lattice) structure, systematically analyzing neighborhood relations and directional connections (Li and Chen, 2019). This enables meaningful segmentation results even in irregular and challenging image patterns.

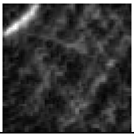
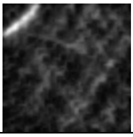
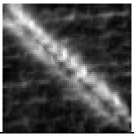
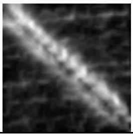
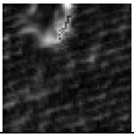
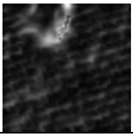
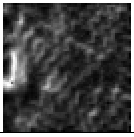
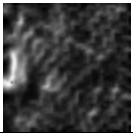
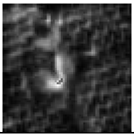
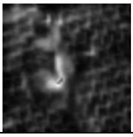
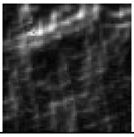
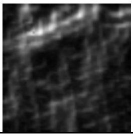
In the literature, lattice segmentation has demonstrated superior performance compared to classical region-based or edge-based segmentation techniques. This advantage stems from the fact that pixel relationships are based not only on intensity similarity but also on structural similarities and directional coherence (Kumar et al., 2020). Thus, micro-level anomalies or defects can be revealed consistently and contextually within the structural framework. The benefits of lattice segmentation have been frequently emphasized in critical applications such as surface defect detection and medical image analysis (Singh and Patel, 2021).

Lattice segmentation commonly employs multiscale analysis and directional filters during implementation. These filters decompose structural information at different orientations within the image, providing a rich feature set to the segmentation algorithm (Wang et al., 2018). Moreover, the lattice structure contributes to regional integrity and consistency in the segmentation process, reducing the influence of noise and minor distortions (Huang and Zhang, 2019).

In conclusion, lattice segmentation successfully delivers accurate segmentation results in complex and low-contrast images by jointly leveraging directional and structural information. Its integration with multiscale preprocessing techniques such as RCT significantly enhances the overall performance of image processing systems.

Figure 2 Applying Lattice Segmentation to RCT Data

RCT data	*Lattice segmentation output*
	
	
	
	

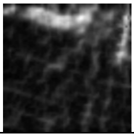
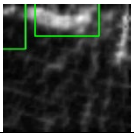
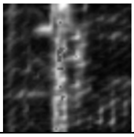
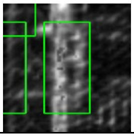
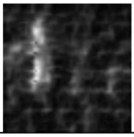
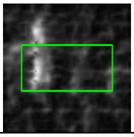
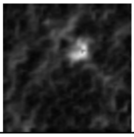
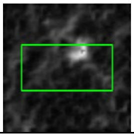
	
	
	
	
	
	

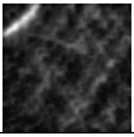
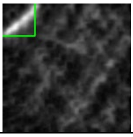
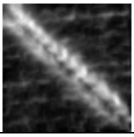
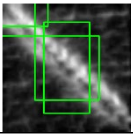
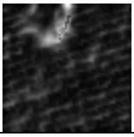
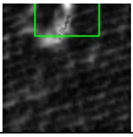
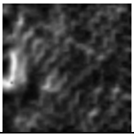
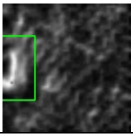
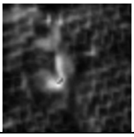
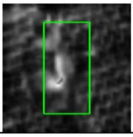
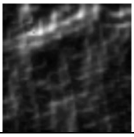
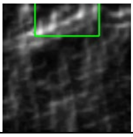
Region Proposal Network (RPN) for Meaningful Region Proposal: Masks obtained after image segmentation are often insufficient for further tasks such as defect or object detection. At this stage, the meaningful region proposal step becomes crucial by automatically identifying candidate regions of interest. The RPN is a deep learning-based architecture developed specifically for this purpose and has become prominent especially in object detection tasks [Ren et al., 2015].

RPN generates candidate regions (region proposals) on the segmentation output using a sliding window approach with varying scales and aspect ratios. These candidate regions provide data for subsequent classification and defect detection models, enabling the grouping of pixels not only individually but also as meaningful cohesive units [Girshick, 2015]. This approach allows for highly precise localization of defect regions, particularly in images containing complex and small surface defects [Ren et al., 2017].

In the literature, RPN applications have demonstrated successful results in fields such as medical imaging [Liu et al., 2019], industrial quality control [Zhang et al., 2020], and textile defect detection. A key advantage of RPN is that by automatically segmenting the data into regions, defect detection models can operate more rapidly and effectively without requiring human intervention [He et al., 2016].

Figure 3 Applying RPN on Lattice Segmentation Outputs

Lattice segmentation data	*Defective Region Proposals Using RPN*
	
	
	
	

Conclusions

In this study, lattice segmentation, RCT, and RPN-based methods were integrated to achieve high-accuracy segmentation and defect detection in images with complex structures. RCT effectively captures multi-directional and multi-scale structural information of images, enabling better representation of fine details and directional features, especially on surfaces. When combined with lattice segmentation, this significantly improved segmentation quality and detail preservation. The RPN algorithm applied after segmentation automatically proposes potential defect regions, accelerating data processing and enhancing detection accuracy. Thus, the proposed integrated approach substantially improved system performance in both segmentation and defect detection stages.

Experimental results demonstrated that RCT-based multi-directional feature extraction provides advantages in detecting small and low-contrast defects, particularly on complex surfaces. The structural modeling power of lattice segmentation combined with the fast object proposal capability of RPN yielded highly accurate and reliable results in industrial quality control and medical imaging applications. Future work will focus on adapting these methods for real-time applications, conducting comprehensive tests on different data types, and investigating integration with deep learning-based classification algorithms. These developments are expected to increase the system’s generalizability and enable its effective use in a broader range of application domains.

References

Chen, Y., Wang, L., & Zhang, Q. (2013). A redundant contourlet transform for image denoising. *Signal Processing*, 93(10), 2857-2870.

Ding, Y., & He, Y. (2020). Rail surface defect detection using attention-guided deep learning. *IEEE Transactions on Industrial Informatics*, 16(5), 3253–3261.

Do, M. N., & Vetterli, M. (2005). The contourlet transform: An efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12), 2091–2106.

Gonzalez, R. C., & Woods, R. E. (2018). Digital Image Processing. Pearson.

Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440–1448.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, M., & Zhang, Q. (2019). Robust lattice segmentation techniques for noisy image environments. *IEEE Access*, 7, 124512-124523.

Kumar, R., Singh, P., and Verma, S. (2020). Lattice structure analysis for surface defect detection in industrial images. *Pattern Recognition Letters*, 133, 244-251.

Li, X., & Chen, J. (2019). Spatial lattice modeling for image segmentation: A survey. *IEEE Transactions on Image Processing*, 28(3), 1401-1415.

Liew, A. W., et al. (2005). Fuzzy image clustering incorporating spatial continuity. *IEEE Transactions on Image Processing*, 15(2), 572–581.

Liu, F., Song, E., and Zhang, W. (2019). Application of Region Proposal Network for Tumor Detection in Medical Images. *IEEE Access*, 7, 100053–100062.

Liu, X., Zhao, Y., & Zhang, M. (2020). Textile surface defect detection based on multiscale directional transform. *Textile Research Journal*, 90(7-8), 787-800.

Mallat, S. (2009). A Wavelet Tour of Signal Processing. Academic Press.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.

Pham, D. L., Xu, C., & Prince, J. L. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1), 315–337.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 91–99.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Object detection networks on convolutional feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7), 1476–1481.

Singh, A., and Patel, D. (2021). Applications of lattice segmentation in medical image analysis: A comprehensive review. *Computer Methods and Programs in Biomedicine*, 202, 106001.

Sun, R., & Tang, Y. (2021). Medical image enhancement using redundant directional multiscale transform. *Computer Methods and Programs in Biomedicine*, 204, 106051.

Wang, J., & Chen, Z. (2019). Multidirectional filtering for surface defect detection in industrial images. *Journal of Manufacturing Systems*, 52, 64-74.

Wang, T., Li, H., and Zhao, J. (2018). Directional filtering and multi-scale analysis in lattice-based image segmentation. *Signal Processing: Image Communication*, 61, 97-108.

Wang, X., et al. (2016). Textile defect detection based on improved deep learning networks. *Textile Research Journal*, 86(12), 1237–1250.

Zhang, H., & Li, J. (2018). Directional multi-scale transform for texture analysis. *Pattern Recognition Letters*, 105, 43-49.

Zhang, Y., et al. (2019). Directional texture analysis for surface defect detection. *Pattern Recognition Letters*, 125, 546–554.

Zhang, Y., Wang, L., and Chen, H. (2020). Automated defect detection in industrial quality control using RPN-based deep learning. *Computers in Industry*, 123, 103307.

Zhao, L., and Wang, Y. (2017). Advanced lattice-based segmentation techniques for complex textures. *Journal of Visual Communication and Image Representation*, 45, 12-25.

CHAPTER 5

The Role of Artificial Intelligence in Genetic Systems

Muhammet MEYDAN[1]

Volkan KAYA[2]

1. Introduction

People learn from the situations they experience since childhood and reflect them in their lives. The emergence of artificial intelligence is also creating computer systems that aim to think like

[1] Master's Student, Erzincan Binali Yıldırım University, Graduate School of Natural and Applied Sciences, Department of Artificial Intelligence and Robotics, Erzincan/Türkiye, Orcid: 0009-0005-4379-9718, muhammet.meydan@ogr.ebyu.edu.tr

[2] Assoc. Prof. Dr., Erzincan Binali Yıldırım University, Faculty of Engineering and Architecture, Department of Computer Engineering, Erzincan/Türkiye, Orcid: 0000-0001-6940-3260, vkaya@erzincan.edu.tr

humans (Chen, 2019). Artificial intelligence is actually an intelligent software. Therefore, the common point between artificial intelligence and humans is accepted as the ability to learn and solve problems (Bozkurt & Armağan, 2020). Intelligence is considered as the ability to gather and combine necessary information to solve a complex problem (Feigenbaum, 1989). The criteria for intelligent behavior in intelligent systems are the ability to perceive, think and act. Thanks to these abilities, incoming signals can be perceived and analyzed (Kocabas, 1991). In order for a system to be considered intelligent, it must have the ability to learn (Alpaydın, 2020). Artificial intelligence is a field of research that aims to analyze human mental functions with computers and adapt them to artificial systems (Russell et al., 1995).

Artificial intelligence was most concretely introduced in 1943 by McCulloch and Pitts (McCulloch & Pitts, 1943) in a study called "Boolean Circuit Model of the Brain". In this study, the operation of neurons in the human brain was explained mathematically. Therefore, the formulation of the working method of the human brain is also considered as one of the most important stages of artificial intelligence.

Artificial intelligence was first brought to the agenda by John McCarthy at the Dortmund Conference in 1956. In addition, this conference in 1956 was the first conference given on artificial intelligence. In this conference, McCarthy defined artificial intelligence as the branch of science and engineering that deals with human-like thinking beings (Sosyal, 2021).

By the 1960s, computers not only could run commands but also had the ability to store data. With this development, different types of studies on artificial intelligence emerged. The first work in this period was the "general problem solver" developed by Newell and Simon and the program called ELIZA developed by Joseph Weizenbaum in MIT's artificial intelligence laboratories. Designed between 1964 and 1966, ELIZA is considered the first work in the field of natural language processing (Weizenbaum, 1966).

The 1980s are considered to be the period when computers gained the ability to establish relationships between information. In this case, computers gained the ability to use the data they had stored or used in the past in new experiences. This development was called "machine learning." During the same period, Edward Feigenbaum developed "expert systems" that could imitate the stages that humans think about when making decisions.

In the 1990s, artificial neural networks were developed. The working method of artificial neural networks emerged as “learning systems” that train themselves with sample data, without drawing a roadmap for a specific task. In other words, if an owl image is given as input to the machine and then randomly entered images are entered, it is defined as being able to guess whether there is an owl image or not.

The first goal of artificial intelligence is to create machines that can think like humans. In 1997, IBM developed a chess-playing program called Deep Blue, a computer system that managed to beat the famous champion Gary Kasparov. In the same period, the "speech recognition program" used in Windows and created by Dragon Systems was introduced (Pirim, 2006).

In 2000, a robot named "Kısmet" was designed that can make human-like facial movements, make voice and body movements very close to humans, and even speak, socialize like a human and learn as it socializes. The aim of the robot is not only to exhibit movements similar to human movements, but also to establish a social bond with what it learns and to design an intelligent robot that learns through the bonds it establishes (Breazeal, 2004).

Today, the areas of use of artificial intelligence have expanded considerably (Elmas, 2018). Artificial intelligence has evolved over the years and has been divided into broad sub-branches and integrated into different fields. There are many sub-branches of artificial intelligence such as machine learning, robotics, artificial neural networks, natural language processing, and image processing. Artificial intelligence covers many different branches of science, such as psychology, genomics, neurotechnology, health, and finance, in addition to computer science (Sosyal, 2021). Some views even describe it as the most important event since the creation of the universe (McCorduck & Cfe, 2004). Figure 1 shows the chronological history of artificial intelligence.

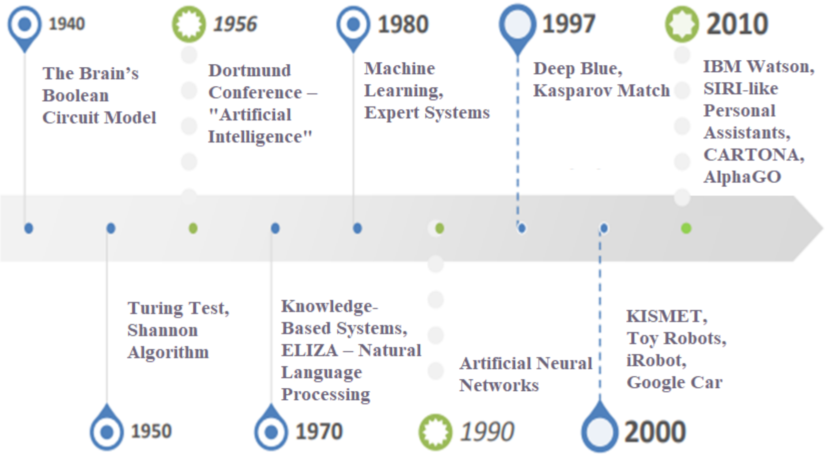

1940
The Brain's Boolean Circuit Model
1956
Dortmund Conference – "Artificial Intelligence"
1980
Machine Learning, Expert Systems
1997
Deep Blue, Kasparov Match
2010
IBM Watson, SIRI-like Personal Assistants, CARTONA, AlphaGO
Turing Test, Shannon Algorithm
1950
Knowledge-Based Systems, ELIZA – Natural Language Processing
1970
Artificial Neural Networks
1990
KISMET, Toy Robots, iRobot, Google Car
2000

Figure 1. Chronological history of artificial intelligence (Arslan, 2020)

Artificial intelligence has integrated into many different fields with its development. One of these fields is the field of health. In the field of health, significant gains have been made with the use of artificial intelligence, especially in studies on genetics.

Genetics is one of the most important branches of science in the 21st century that studies the structure and functions of genes and the inheritance of genetic traits. Genetic factors affect an individual's physical characteristics, behaviors, the possibility of contracting a disease, and many other characteristics (Cho et al., 2009). With the advancement of technology, approaches to solving genetic diseases have also changed and developed. Genetics are seen to be effective in diseases that are common in society, from hypertension to psychological disorders. Genetic effects have come to the fore in the diagnosis, treatment and prevention of disease, and thus the importance of genetics has increased in the medical world (Watson et al., 1999; Menasha et al., 2000). Genetic factors constitute a significant portion of the diseases seen in societies (Çakır, 2020). Nowadays, artificial intelligence is widely used in the field of medicine. Artificial intelligence techniques are frequently encountered, especially in studies on genetics. Significant progress has been made with artificial intelligence in the detailed examination of genetic factors and the extraction of DNA characteristics in the detection of diseases affecting health (Sosyal, 2021).

1.1. Related Works

In this section, studies in the literature in recent years on the use of artificial intelligence in the field of genetics are examined.

Researchers at the Houston Methodist Research Institute in Texas have developed an artificial intelligence software that can accurately predict breast cancer risk by analyzing millions of records from patients' mammogram results. In the study, breast cancer mammography images and pathology results of 500 patients were used as data. According to the results obtained, it was observed that the artificial intelligence software diagnosed the disease with 99% accuracy and 30 times faster than a physician (Griffiths, 2016).

In a study conducted by Er and colleagues, an artificial immune system was used to diagnose lung membrane cancer. Thanks to this algorithm used, diagnoses made using the same disease and data set were compared with artificial neural network results and a 97.74% success rate was achieved. Therefore, in this study, a system has been developed that can significantly assist physicians in diagnosing the disease (Er, 2015).

In the study conducted by Quang and his colleagues, three different methods, namely logistic regression, support vector machine and artificial neural networks, were considered and compared in order to explain the pathology of genetic variants. As a result of the comparison, the classification success accuracies of the methods were obtained as 58.2%, 59.8% and 66%, respectively (Quang et al., 2014).

In their study, Suk and colleagues proposed a new method for the representation of high-level latent and shared features with neuroimaging methods using deep learning methods. In this method, Deep Boltzmann Machine was used. To test the effect of the results, experiments were conducted with Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and the obtained results were compared with the state-of-the-art methods. In the study, three different binary classification problems were examined between Alzheimer's patients and healthy individuals, between individuals with mild cognitive impairment and healthy individuals, and between those with progressive and non-progressive mild cognitive impairment. It was observed that the developed model gave more successful results in visual inspections compared to other methods; the model reached 95.35%, 85.67% and 74.58% accuracy rates in three different classification problems, respectively (Suk et al., 2014).

Pereira et al. used a convolutional neural network to analyze handwritten images for computer-aided Parkinson's disease diagnosis. To support this work, a special dataset was prepared that included visual and signal-based data. The proposed method exhibited superior performance with an accuracy rate of up to 95% in early-stage detection when compared to approaches with raw data and texture-based features (Pereira et al., 2018).

In the study conducted by Malkoçoğlu, deep learning, machine learning, transfer learning and convolutional neural network based models were used on digital pathology data of the disease in the classification of Acute Lymphoblastic Leukemia cells. As a result of the classification, it was reported that the Convolutional Neural Network model gave 87% more successful results compared to other models (Malkoçoğlu & Berika, 2020).

2. Genomic Data and Big Data

Big data is a term that has been used a lot in societies in recent years. Especially since the early 2000s, it has been seen that all the characteristics of people (physiological, biological, psychological, etc.) have been converted into data in a digital environment. In this way, data has become more easily accessible and has also been converted into a format that computers can understand (Auffray et al., 2016).

When all biological data is collected and recorded regularly, this information becomes a part of the biological system. As a result, artificial intelligence enables humanization or the mechanization of humans and the medical creation of humans through reproduction with medical assistance or interventions on genes. In this way, it is possible to process genes that provide important data on many diseases (Costa, 2014).

2.1. Genomic Data

Genomics is a branch of science that identifies the entire genetic structure of an organism, examines the interactions of these genes among themselves and with environmental factors, transfers the obtained data to digital media, analyzes it and ensures the storage of this information (Yarman et al., 2003). Genomics enables the comparison of genetic information belonging to different living things, enables the investigation of evolutionary similarities and makes it possible to obtain information about the number, diversity and functions of proteins produced by living things (Şahin-Çevik, 2005).

Thanks to genomic analysis, the genetic structures of individuals can be examined in depth. It also plays an important role in examining their relationships with diseases (Libbrecht et al., 2015). Genetic information is not easy to interpret due to its complex structure (Zou et al., 2019). It is not possible to obtain efficient results using traditional analysis methods with these complex structures. For this reason, it has become necessary to develop new approaches that have the ability to process and interpret high-dimensional data (LeCun et al., 2015).

2.2. The Role of Big Data in Genetic Systems

The concept of “big data” was first mentioned in an article by Michael Cox and David Ellsworth in 1997, in the context of data processing challenges faced by computer technologies (Cox & Ellsworth, 1997). There are criteria that data must meet to be defined as large. These features are briefly called 5V features (Gürsakal, 2014). 5V features;

- Volume: The space occupied by the data in the system
- Velocity: The continuous growth of the data set with its ongoing production
- Variety: The diversification of the available data with new data sources
- Verification: The security of big data
- Value: It is expressed as the material value that the data will provide (Ffoulkes, 2017).

The efficient processing of big data contained in genomic information by artificial intelligence algorithms has increased the speed and accuracy of research in this context, allowing only sensitive analyses to be performed. It has also played a major role in revealing biological meanings that have not yet been discovered (Zou et al., 2019). With the processing of big data, diseases can be diagnosed early. At the same time, genetic variants can be classified clinically and personalized treatment methods can be applied. In this way, important contributions have been made to the creation of personalized medical applications in modern medicine (Libbrecht & Noble, 2015).

2.3. Processing of Data with Artificial Intelligence

Genetic studies can reveal biomarkers that can detect various diseases, predict risk, and predict treatment outcomes (Zeeshan et al., 2020). Most genetic studies investigate biological predictions and disease risks by comparing healthy and patient populations, which may miss individual and subgroup variations (Ahmed et al., 2020). DNA and RNA sequencing are the most commonly used methods in genetic research.

Genetic variations, which describe DNA and RNA differences, are an important element for understanding the genetic basis of disease (Martin et al., 2021). DNA and RNA sequencing can determine the association between diseases and genomic variants (Ahmed et al., 2021; Ahmed et al., 2021). Low and high polygenic scores obtained from DNA can give the probability of developing the disease (Lewis & Vassos, 2020). Although all these results are promising, the challenge here is to analyze big data and use the analysis results in a way that physicians can understand to diagnose, determine risk, and predict treatment outcomes (Ahmed et al., 2021).

In this direction, artificial intelligence is used to use a large genetic pool to analyze disease results accurately and safely

(Abdelhalim et al., 2022). The collection of genomic data and processing with artificial intelligence has great potential (Ahmed et al., 2020). The applied artificial intelligence methods try to perform the learning process from a dataset that shows a lot of details (Ahmed, 2021). Artificial intelligence approaches offer statistical analysis of genomic data to determine the best sources of information and predict high-risk diseases (Ahmed et al., 2020).

3. Use of Artificial Intelligence Techniques in Genetic Systems

Artificial intelligence is a system that aims to think like the human brain by imitating human intelligence. While performing these operations, it is divided into fields in order to obtain better results against the conditions in different fields and approaches. In this section, the effects of these sub-branches on genetic systems and their usage examples are discussed.

3.1. Machine Learning Methods

Machine learning can be considered as the logic of learning from human experience, adapted to computers. It is a sub-branch of artificial intelligence that allows computers to learn through experience.

A large number of machine learning methods have been developed for the use of machine learning in genetic systems and to understand gene expression in detail. Some methods aim to predict the expression of a gene from the DNA sequence alone (Beer &

Tavazoie, 2004), while others take into account ChIP-seq histone modification (Karlić, 2010) or Transcription Factor (TF) binding (Ouyang et al., 2009) profiles in the gene promoter region. More advanced methods attempt to jointly model the expression of all genes in a cell by training a network model (Friedman, 2004). A study demonstrates the methods used in 24 different peer-reviewed and published scientific studies that address machine learning algorithms for the analysis of genomic data (Figure 2). The studies cover shared genomic data for various diseases.

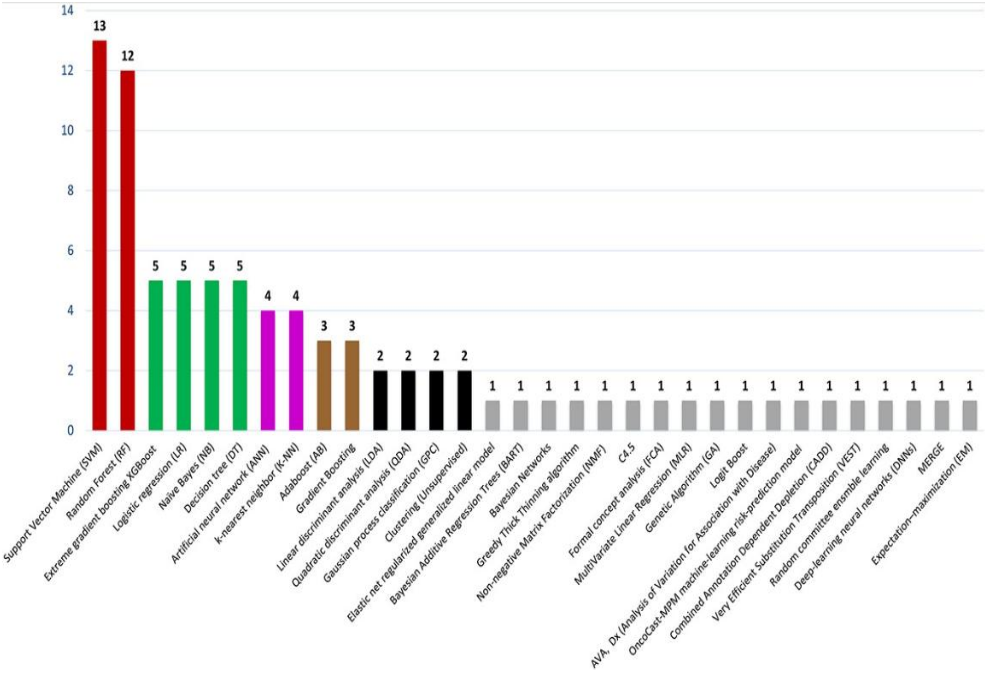


Figure 2. Machine learning algorithms applied for analysis of genomic data (Vadapalli et al., 2022)

3.2. Deep Learning Methods

Deep learning is considered a subfield of machine learning and focuses on extracting meaning from complex data, especially using multi-layered artificial neural networks. Deep learning can analyze large data sets in a short time with artificial neural network models. Artificial neural networks are mathematical models designed by taking inspiration from neuron structures in the human brain (LeCun et al., 2015). It can be said that imaging data in particular is quite large and complex. Deep learning can achieve a high success rate in such data. This method is frequently used in fields such as radiology and pathology.

With the rapid accumulation of large-scale data such as genetic data in the last decade, deep learning has attracted intense attention (Hanbay, 2019; Hatipoğlu & Altuntaş, 2024). It has shown superior performance compared to machine learning approaches in many areas such as sequence motif detection (Hatipoğlu & Altuntaş, 2024; Lanchantin et al., 2017), chromatin interaction prediction (Singh et al., 2019; Whalen et al., 2016; Zeng et al., 2018) and genetic variant detection (Zhou & Troyanskaya, 2015).

In this chapter, deep learning methods that are frequently used in genetic and genomic research are reviewed. Additionally, problems such as DNA/RNA binding sequence motif identification (Hatipoğlu & Altuntaş, 2024; Lanchantin et al., 2017; Zhou & Troyanskaya, 2015 ; Kelley et al., 2016; Quang & Xie, 2016) and gene expression prediction (Hanbay, 2019; Zhou & Troyanskaya, 2015; Kelley et al., 2016; Quang & Xie, 2016; Singh et al., 2016) are addressed. Five basic methods are generally used in such studies: input modification, input reconstruction, saliency maps, convolution kernel analysis, and attention mechanisms (Figure 3).

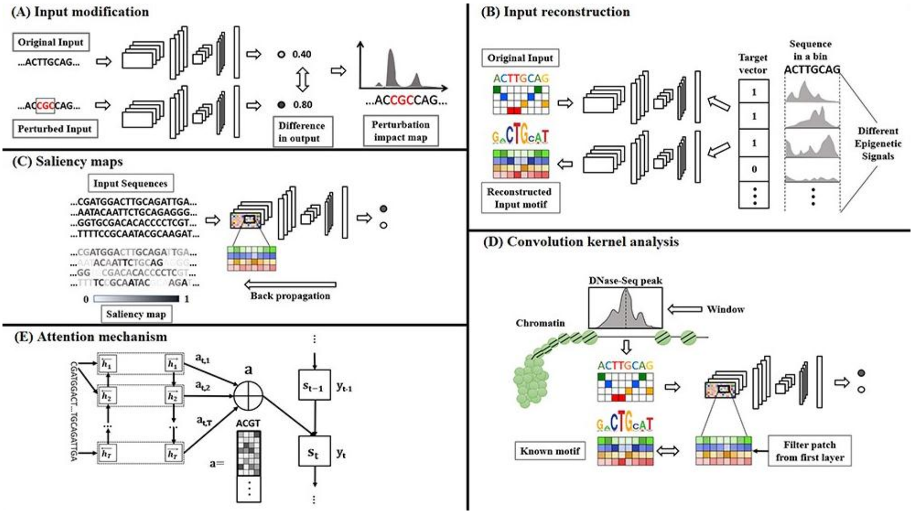


Figure 3. Deep learning methods used in genetics and genomics (Talukder et al., 2021)

Figure 3 visually explains deep learning methods used in genetic and genomic research. In Figure 3a, with the input modification method, the effect of changes in the input data on the classification is analyzed and these effects are visualized as a distortion map. In Figure 3b, in the input reconstruction method, the most suitable examples are created in line with the weights of the model and the performance of the model is evaluated by comparing these examples with the original input. In Figure 3c, with the saliency map method, the contributions of the input sequences to the classification are measured and these contributions are visualized using the backpropagation algorithm. In Figure 3d, in the motif discovery method, the first convolution filter of the convolutional neural networks (CNN) is visualized and their biological significance is revealed by comparing with known motifs. Finally, in Figure 3e, with the attention mechanism method, a recurrent neural network (RNN) identifies the important parts of a given input sequence and with this information, an output sequence is generated and a mapping is made to the input sequence.

3.3. Evaluation of Genetic Data with Image Processing

Surveillance and image recognition are the most important application areas of computer vision (Retson et al., 2019).

Computer vision is frequently used in clinical genomic testing. For example, deep learning can identify cancer cells in lung cancer images, determine their type, and predict which somatic mutations are present in the tumor (Rios Velazquez et al., 2017; Coudray et al., 2018). Similarly, facial image recognition can be used to identify rare genetic disorders and guide molecular diagnoses (Gurovich et al., 2019). Figure 4 shows an example of the use of AI in clinical and genomic diagnostics.

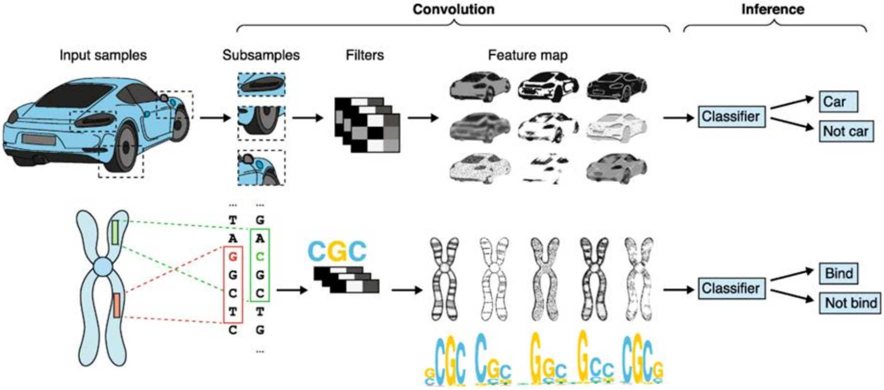


Figure 4. Artificial Intelligence in Clinical and Genomic Diagnostics (Dias & Torkamani, 2019)

Figure 4, convolutional neural networks (CNN) process the input image data (upper part) or DNA sequence (lower part) by dividing it into a multi-layered structure. Thanks to this multi-layered structure, the number of samples is increased by fragmenting the data and various filters are applied to each sub-sample. The features carried by each split sample are determined by these applied filters and these features are multiplied by weights to create new patterns. These new patterns are compared with the existing patterns in the input data to estimate the accuracy of the system. This process emphasizes the ability of AI-supported diagnostics to detect patterns in both clinical and genomic data.

3.4. Natural Language Processing Methods

Natural language processing (NLP) is simply the computational extraction of meaning from human language. Natural language processing takes a document or automatic speech recognition as input and produces a useful transformation of the document as output. This transformation can be language translation, document classification, or summary extraction (Wu et al., 2016). NLP can be used to make health information more accessible by translating educational materials into other languages or translating medical terms into their definitions (Chen et al., 2018). Chatbots with artificial intelligence infrastructure are used to support genetic counselors in meeting demands such as clinical or consumer genetic testing (Kohut et al., 2019). NLP, when combined with genomic data, has been used for rare disease diagnosis and genetic analysis, and has achieved an accuracy rate in genetic diagnoses similar to that of physicians (Liang et al., 2019; Clark et al., 2019). Figure 5 shows the recurrent neural networks (RNN) model used in natural language processing (NLP) applications in clinical and genomic diagnostics.

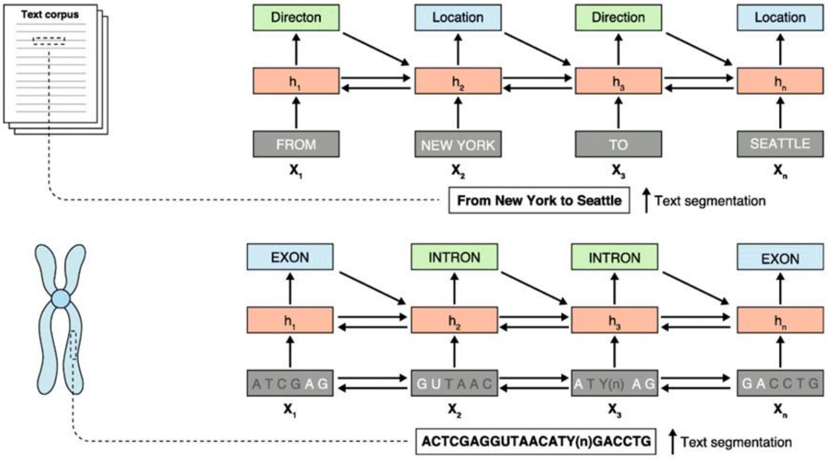


Figure 5. Artificial Intelligence in Clinical and Genomic Diagnostics (Dias & Torkamani, 2019)

Figure 5 shows an artificial neural network model that models relationships between input data (X) via hidden layers (h), working on a text-based piece of data (top) or a DNA sequence (bottom). Hidden layers typically consist of one-way recursive nodes that analyze the data and only pass information forward. However, the example in Figure 5 shows a bidirectional RNN structure that can capture both past (backward) and future (forward) context information. This structure enables more effective learning and interpretation of complex relationships, especially in biological sequences.

4. Artificial Intelligence in Diagnosis and Treatment of Diseases

With the widespread use of artificial intelligence in the health sector, the diagnosis and treatment times of diseases have been significantly improved (Yeasmin, 2019).

It has even been developed to a level where it can detect a person's potential to catch a disease in advance. The study of creating a medical archive by storing patients' medical records is widely used. However, as the size of the data increases, it becomes more complex than it can be processed. With artificial intelligence, the medical data of patients is examined, the diseases that the person has had are analyzed, and the risk of catching future diseases becomes predictable.

4.1. Early Diagnosis of Genetic Diseases

Artificial intelligence plays an important role in the early diagnosis of genetic diseases thanks to its ability to analyze genetic data. Genetic data is quite complex and large. Analyzing such a large amount of data using traditional methods is in a very weak position in terms of time and successful results. Artificial intelligence can analyze this data very quickly and with a high level of accuracy (Hanbay, 2019).

4.2. Artificial Intelligence in the Development of New Treatment Methods

With the use of artificial intelligence in the medical field, significant innovations have emerged in clinical processes. Thanks to these technological developments, physicians' observation and diagnosis processes have been strengthened with artificial intelligence-supported systems, and artificial intelligence applications have taken their place as an additional auxiliary tool in decision-making processes. An example of the contributions of artificial intelligence to the health field is Enlitic, a company operating in the field of medical imaging. Enlitic develops systems that support physicians in the analysis of imaging data using deep learning-based algorithms.

Enlitic contributes to clinical decision support systems by analyzing multi-dimensional clinical data such as radiological and pathological images, ECG records, blood tests, genomic data, patient history and electronic health records through deep learning-based algorithms. In this way, both the accuracy rate in diagnostic processes is increased and a more comprehensive evaluation of diseases becomes possible. The ability of artificial neural networks to analyze large-scale data without human intervention allows the models developed by Enlitic to learn on millions of medical images. Large-scale medical image analyses, which are very difficult for physicians to perform under current conditions, can be performed by artificial intelligence systems in milliseconds. This means that the interpretation speed is approximately 10,000 times faster than that of a physician. In addition, comparative analyses show that artificial intelligence systems produce results with higher accuracy rates than physicians' determinations in many cases. For example, it has been reported that the system developed by Enlitic provides approximately 50% more accurate and faster results compared to the evaluations of three different radiologists in the classification of malignant tumors (Uzun, 2020).

It can be said that one of the biggest effects of the success rate in the treatment of genetic diseases with artificial intelligence is that artificial intelligence detects diseases faster than doctors. This can mean that many new methods developed perform better than doctors.

4.3. Personalized Medicine and Artificial Intelligence

Despite advances in biomedical medicine, one of the biggest problems seen in recent years is that some patients do not respond positively to drug treatments. According to a report published by the US Food and Drug Administration, drug treatments are found to be ineffective in 38%-75% of people with widespread disease. This situation can cause problems such as the patient's treatment going negatively, as well as suffering and high costs. The reason why people with the same disease do not respond positively to the same type of drugs is due to differences in genes. In the field of digital and genomic medicine, large data sets obtained from a wide variety of sources such as wearable devices, medical images, and electronic health records are analyzed with artificial intelligence algorithms, enabling the development of individualized treatment strategies (Björnsson et al., 2020). There are many examples of treatments in personalized medicine, almost all of which are based on individuals' genetic profiles. In other words, drugs suitable for each patient's genetic structure will be developed, the treatment process can be carried out using these drugs, and these data can then be used for early diagnosis of the disease (Goetz & Schork, 2018).

5. Challenges and Advantages of Using Artificial Intelligence in Genetic Systems

Big data analysis is a very time-consuming process, and humans perform such operations much longer than machines.

Furthermore, it is often not possible for several people working together to complete a task at the same speed that a machine can do alone. In this context, one of the biggest advantages offered by artificial intelligence is that it can produce faster and more efficient solutions. This speed provides a great benefit, especially in terms of early diagnosis and intervention. Early diagnosis offers an important opportunity for faster detection and treatment of diseases, while also allowing treatment processes to be completed at lower costs (Tarcan et al., 2024).

Another positive contribution of artificial intelligence in genetic systems is its impact on the drug development process. Information obtained from the patient's personal data can be analyzed by artificial intelligence algorithms to create personalized treatment strategies. This process helps develop personalized medicine and develop more effective and targeted treatment methods (Björnsson et al., 2020).

In addition, the effects of artificial intelligence in the field of health care have significantly reduced the workload of healthcare personnel and increased their efficiency (Mesko, 2017). However, the use of artificial intelligence in genetic systems also brings with it some challenges. Data sharing, security measures and ethical issues are among the obstacles to the wider application of this technology. In addition, the issue of responsibility for errors that may occur in artificial intelligence-based systems is still an area of uncertainty and this is an important topic of discussion.

5.1. Technical and Scientific Challenges Encountered

There are many positive aspects of artificial intelligence in genetics and health systems. However, in the event of a clinical error that may occur in the application of these systems, who will be held responsible is still a matter of debate. Today, the decisions made by artificial intelligence are evaluated by medical professionals, and the final decision-making authority belongs to these experts. However, it is thought that in the future, artificial intelligence has the potential to make more independent decisions and there is a risk of implementing these decisions. It is predicted that this situation may have serious consequences in health practice (Mesko, 2017).

Although artificial intelligence is designed to be similar to human thinking, sometimes it may not fully reflect human-like thinking. In particular, some clinical situations have led to studies questioning the decision-making ability of artificial intelligence. As an example, studies on skin cancer discuss the existence of lesions that cannot be seen with the naked eye. Such situations may require more detailed examination and expert evaluation. In this context, it is argued that artificial intelligence should be used as a supporting tool for clinical decisions rather than as a decision maker (Mar & Soyer, 2018).

5.2. Advantages and Disadvantages of Using Artificial Intelligence in Genetic Systems

The use of artificial intelligence technologies in genetic systems offers great advantages in terms of both accelerating diagnostic processes and developing personalized treatment opportunities. However, the implementation of these technologies also brings with it some ethical, security and data sharing issues. The main advantages and disadvantages of using artificial intelligence in genetic systems are summarized below:

Advantages:

- It saves time and cost in genome sequencing processes.
- It increases analysis accuracy by minimizing human error.
- It quickly identifies genetic variants and diseases.
- It increases treatment effectiveness by supporting personalized drug development processes.
- It can process large and complex data with high accuracy.

Disadvantages:

- Requires high security measures in terms of protecting personal data.
- Reluctance to share data among private institutions makes cooperation difficult.

- Determining ethical responsibility and legal liability for errors caused by artificial intelligence is unclear.

5.3. The Future of Artificial Intelligence Applications in Genetics

The role of artificial intelligence in genetic systems is increasing day by day and it is becoming possible to make various predictions about the future. Of course, time will show to what extent these predictions will come true. However, considering current developments, it seems likely that certain changes will occur in the short, medium and long term.

In the near future, with the integration of artificial intelligence into genetic systems, it is expected that traditional diagnosis and treatment methods will be replaced by faster and more effective AI-supported applications. With the more widespread use of personalized medicine, patient-specific treatment approaches will be developed and diagnostic processes will become more reliable. In addition, it will be possible to significantly speed up time-consuming processes such as processing and analyzing genetic data through automation.

In the medium term, the diagnosis and treatment of genetic diseases may become simpler, more accessible, and less costly. With the contribution of artificial intelligence in drug development processes, personalized treatment protocols can be applied more effectively. In addition, thanks to artificial intelligence-supported health applications integrated with mobile devices, the diagnosis of some diseases may reach a level where it can be done with just one imaging or brief analysis.

In the long term, AI-based robotic systems may assume more responsibility in the healthcare sector. It may be possible to develop systems where diagnosis and treatment processes are completely carried out by AI and do not require human intervention. In this context, the nature of healthcare services may change and especially repetitive or standardized processes may be transferred to AI-supported systems. Of course, the extent to which these processes will be implemented will depend on ethical, legal and social factors as well as technological advances.

As a result, the integration of AI and genetic systems could lead to revolutionary developments in the field of healthcare in the future. Current technological trends suggest that much more comprehensive AI applications will be witnessed in the coming years.

6. Conclusions and Evaluation

The effect of artificial intelligence on genetic systems has produced many positive results. People are freed from many unnecessary elements with artificial intelligence-supported diagnosis and treatments. The time and money lost due to wrong treatment methods have decreased significantly with the use of artificial intelligence. In addition to these losses, artificial intelligence seems to be used in genetic systems as a new and promising method for many patients who do not respond positively to treatment. Unnecessary and incorrect medication use, which is one of the most common problems encountered during treatment, seems to be a factor that can reverse the positive course of treatment. Personalized medication use developed with artificial intelligence positively affects patients both financially and in terms of regaining their health. Artificial intelligence not only helps patients, but also facilitates the work of doctors to the extent that it can be called their right arm. Although analyzing large and complex data is difficult for a human, this difficulty has been largely eliminated with the use of artificial intelligence, and doctors have become able to easily and quickly interpret the analysis of the data they want. The greatest contribution of artificial intelligence to genetic systems is seen in the direction of fast and accurate detection and prediction of diseases that will be encountered in the future. Although knowing in advance which disease you are at risk of contracting five years from now seems unrealistic at first glance, with the development of artificial intelligence technologies, this has now become possible.

6.1. Potential Future Impacts

Artificial intelligence is widely used today to assist physicians. The final decision on the data obtained is made by physicians. In the near future, robot doctors in particular may be designed to make their own decisions regarding treatment. Today, treatments for diseases with a high risk of death will yield more positive results. In fact, diseases such as cancer can be diagnosed years before the disease is contracted. In addition, personalized medicine, which will be very effective in preventing unnecessary drug use and speeding up treatment, will become widespread. The era of using drugs according to the patient, not the disease, will come.

6.2. The Importance of Interdisciplinary Collaboration

In order to implement all these developments, it is not enough to just analyze the data. It is of great importance that the analysis results obtained are evaluated by competent physicians and their accuracy is confirmed. Although artificial intelligence has begun to simplify complex structures in genetic systems, it should not be forgotten that the final results must undergo expert supervision. In this context, the correct interpretation of the obtained data is as critical a process as the sharing of accurate data.

References

Abdelhalim, H., Berber, A., Lodi, M., Jain, R., Nair, A., Pappu, A., ... & Ahmed, Z. (2022). Artificial intelligence, healthcare, clinical genomics, and pharmacogenomics approaches in precision medicine. *Frontiers in genetics*, *13*, 929736

Ahmed, Z. (2021). Intelligent health system for the investigation of consenting COVID-19 patients and precision medicine. *Personalized medicine*, *18*(6), 573-582

Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, 1–35.

Ahmed, Z., Renart, E. G., & Zeeshan, S. (2021). Genomics pipelines to investigate susceptibility in whole genome and exome sequenced data for variant discovery, annotation, prediction and genotyping. *PeerJ*, *9*, e11724.

Ahmed, Z., Renart, E. G., Zeeshan, S., & Dong, X. (2021). Advancing clinical genomics and precision medicine with GVViZ: FAIR bioinformatics platform for variable gene-disease annotation, visualization, and expression analysis. *Human genomics*, *15*(1), 37.

Ahmed, Z., Zeeshan, S., Mendhe, D., & Dong, X. (2020). Human gene and disease associations for clinical-genomics and precision medicine research. *Clinical and translational medicine*, *10*(1), 297-318.

Alpaydın, E. (2020) Yapay Öğrenme: Yeni Yapay Zekâ, Çev: Aylin Ağar, *Tellekt, İstanbul.*

Arslan, K. (2020). Eğitimde yapay zekâ ve uygulamaları. *Batı Anadolu Eğitim Bilimleri Dergisi*, *11*(1), 71-88.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., ... & Zanetti, G. (2016). Making sense of big data in health research: towards an EU action plan. *Genome medicine*, *8*, 1-13.

Beer, M. A., & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, *117*(2), 185-198.

Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., ... & Swedish Digital Twin Consortium. (2020). Digital twins to personalize medicine. *Genome medicine*, *12*, 1-4

Bozkurt, Y. & Armağan, E. (2020). Buluşçu Yapay Zekâ ve Patent Hukuku. *Aristo Yayınevi, İstanbul*.

Breazeal, C. L. (2004). *Designing sociable robots*. MIT Press.

Chen, J., Druhl, E., Polepalli Ramesh, B., Houston, T. K., Brandt, C. A., Zulman, D. M., ... & Yu, H. (2018). A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research*, *20*(1), e26.

Chen, L. P. (2019). Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar: Foundations of machine learning: The MIT Press, Cambridge, MA, 2018, 504, 1793–1795.

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H. J., ... & Kim, H. L. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature genetics*, *41*(5), 527-534.

Clark, M. M., Hildreth, A., Batalov, S., Ding, Y., Chowdhury, S., Watkins, K., ... & Kingsmore, S. F. (2019). Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science translational medicine*, *11*(489), eaat6177.

Costa, F. F. (2014). Big data in biomedicine. *Drug discovery today*, *19*(4), 433-440.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., ... & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, *24*(10), 1559-1567.

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In *Proceedings. Visualization'97*, 235-244.

Çakır, D. E. (2020). Nuriye.“. *Genetik Geçişli Hastalıklar, Akraba Evliliği ve Prekonsepsıyonel Bakım, Danışmanlık”(173-197). Prekonsepsıyonel Bakım ve Danışmanlık. ed. Gülbahtiyar Demierel-Fatma Deniz Sayıner. Ankara: Akademisyen Kitabevi*.

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome medicine*, *11*(1), 70.

Elmas, Ç. (2018). *Yapay zeka uygulamaları*. Ankara: Seçkin Yayıncılık.

Er, O., Tanrikulu, A. Ç., & Abakay, A. (2015). Use of artificial intelligence techniques for diagnosis of malignant pleural mesothelioma. *Dicle Medical Journal*, *42*(1), 5-11.

Feigenbaum, E. (1989). Interviewed for expert systems by Kenneth Owen. *Expert Systems*, *6*(2), 112-115.

Ffoulkes, P. (2017). InsideBIGDATA guide to the intelligent use of big data on an industrial scale. *InsideBIGDATA, Massachusetts*.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, *303*(5659), 799-805.

Goetz, LH & Schork, NJ. (2018). Personalized medicine: motivation, challenges, and progress. Fertility and infertility, 109 (6), 952-963

Griffiths, S. (2016). This AI Software Can Tell If You're at Risk from Cancer before Symptoms Appear'. *Şubat*, *23*, 2019.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., ... & Gripp, K. W. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, *25*(1), 60-64.

Gürsakal, N. (2014). Büyük veri. *Baskı, Bursa: Dora Kitapevi*.

Hanbay, K. (2019). Evrişimsel sinir ağı ve iki-boyutlu karmaşık gabor dönüşümü kullanılarak hiperspektral görüntü sınıflandırma. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi,* 35(1), 443-456.

Hatipoğlu, A., & Altuntaş, V. (2024) DeepTFBS: Transkripsiyon Faktörü Bağlanma Bölgeleri Tahmini İçin Derin Öğrenme Yöntemleri Kullanan Hibrit Bir Model. *Politeknik Dergisi*, 1-1.

Karlić, R., Chung, H. R., Lasserre, J., Vlahoviček, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, *107*(7), 2926-2931.

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7), 990-999

Kocabas, S. (1991). Conflict resolution as discovery in particle physics. *Machine Learning*, *6*, 277-309.

Kohut, K., Limb, S., & Crawford, G. (2019). The changing role of the genetic counsellor in the genomics era. *Current Genetic Medicine Reports*, *7*, 75-84.

Lanchantin, J., Singh, R., Wang, B., & Qi, Y. (2017). Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. In *Pacific symposium on biocomputing 2017* (pp. 254-265).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, *12*(1), 44.

Liang, H., Tsui, B. Y., Ni, H., Valentim, C. C., Baxter, S. L., Liu, G., ... & Xia, H. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine*, *25*(3), 433-438.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321-332.

Malkoçoğlu, V., Berika, A. (2020). *Akut lenfoblastik lösemi hücrelerinin derin öğrenme yöntemleri ile sınıflandırılması* (Master's thesis, Lisansüstü Eğitim Enstitüsü).

Mar, V. J., & Soyer, H. P. (2018). Artificial intelligence for melanoma diagnosis: how can we deliver on the promise?. *Annals of Oncology*, *29*(8), 1625-1628

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2021). Publisher correction: clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, *53*(5), 763-763.

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115-133.

Menasha, J. D., Schechter, C., & Willner, J. (2000). Genetic testing: a physician's perspective. *The Mount Sinai journal of medicine, New York*, *67*(2), 144-151.

Mesko, B. (2017). Yapay Zekayla Tıbbi Karar Almak. B. Mesko içinde, Tıbbın Geleceğine Yolculuk (s. 174-183). *İstanbul: Optimist Yayın Grubu*.

Ouyang, Z., Zhou, Q., & Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, *106*(51), 21521-21526.

Pereira, C. R., Pereira, D. R., Rosa, G. H., Albuquerque, V. H., Weber, S. A., Hook, C., & Papa, J. P. (2018). Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification. *Artificial intelligence in medicine*, *87*, 67-77.

Pirim, A. G. H. (2006). Yapay zeka. *Yaşar Üniversitesi E-Dergisi*, *1*(1), 81-93.

Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, *44*(11), e107-e107.

Quang, D., Chen, Y., & Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, *31*(5), 761-763.

Retson, T. A., Besser, A. H., Sall, S., Golden, D., & Hsiao, A. (2019). Machine learning and deep neural networks in thoracic and cardiovascular imaging. *Journal of thoracic imaging*, *34*(3), 192-201

Rios Velazquez, E., Parmar, C., Liu, Y., Coroller, T. P., Cruz, G., Stringfield, O., ... & Aerts, H. J. (2017). Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer research*, *77*(14), 3922-3930.

Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, *25*(27), 79-80.

Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, *32*(17), i639-i648.

Singh, S., Yang, Y., Póczos, B., & Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, *7*(2), 122-137.

Soysal, T. (2021). Crispr Genom Düzenleme Teknolojileri: Patentlenebilirlikleri ve Covid-19 Salgınında Kullanımı. *Adalet Dergisi*, (66), 227-292.

Suk, H. I., Lee, S. W., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, *101*, 569-582.

Şahin-Çevik, M. (2005). Mikroarray Teknolojisi ve Bitkilerde Uygulama Alanları. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *9*(3).

Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, *22*(3), bbaa177.

Tarcan, G. Y., Balçık, P. Y., & Sebik, N. B. (2024). Türkiye ve dünyada sağlık hizmetlerinde yapay zekâ. *Mersin Üniversitesi*

Tıp Fakültesi Lokman Hekim Tıp Tarihi ve Folklorik Tıp Dergisi, *14*(1), 50-60.

Vadapalli, S., Abdelhalim, H., Zeeshan, S., & Ahmed, Z. (2022). Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Briefings in bioinformatics*, *23*(5), bbac191.

Watson, E. K., Shickle, D., Qureshi, N., Emery, J., & Austoker, J. (1999). The 'new genetics' and primary care: GPs' views on their role and their educational needs. *Family Practice*, *16*(4), 420-425.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Whalen, S., Truty, R. M., & Pollard, K. S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics*, *48*(5), 488-496.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yarman, B. S., Gürkan, H., Güz, Ü., & Aygün, B. (2003). A new modeling method of the ECG signals based on the use of an optimized predefined functional database. *Acta Cardiologica-An International Journal of Cardiology*, *58*(3), 59-61.

Yeasmin, S. (2019). Benefits of artificial intelligence in medicine. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1-6). IEEE.

Zeeshan, S., Xiong, R., Liang, B. T., & Ahmed, Z. (2020). 100 years of evolving gene–disease complexities and scientific debutants. *Briefings in bioinformatics*, *21*(3), 885-905.

Zeng, W., Wu, M., & Jiang, R. (2018). Prediction of enhancer-promoter interactions via natural language processing. BMC genomics , 19 , 13-22.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, *12*(10), 931-934.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature genetics*, *51*(1), 12-18.

CHAPTER 6

FROM TIME-FREQUENCY FEATURES TO SMART DIAGNOSİS: A DEEP LEARNING APPROACH FOR BEARING FAULTS

CANAN TASTIMUR[1]

Introduction

The reliability of industrial machinery is of critical importance for sustainable production processes and cost-effective maintenance strategies. Bearings, as fundamental components of rotating machinery, play a vital role in determining the overall health of the system. Early fault diagnosis in bearings is crucial to prevent unexpected downtime and reduce maintenance costs. Therefore, bearing fault diagnosis has attracted significant attention from researchers in recent years, leading to the development of many traditional and deep learning-based methods (Kim and Park, 2020).

Traditional methods often operate directly on raw signals using statistical or frequency-based analyses, whereas nowadays, extracting meaningful features from signals and evaluating these features with machine learning or deep learning models yield more effective results (Wang et al., 2021). Time-frequency based methods

[1] Assistant Professor, Erzincan Binali Yıldırım University, Department of Computer Engineering, Orcid: 0000-0002-3714-6826

capture the characteristics of bearing vibration signals in both time and frequency domains, providing richer features. In this context, methods such as Wavelet Transform, Envelope Analysis, and Short-Time Fourier Transform (STFT) have gained prominence (Zhou et al., 2019).

However, strong feature extraction alone is not sufficient. Appropriate modeling techniques are needed to effectively process these features. In this study, features extracted via time-frequency analysis methods are fed into a Long Short-Term Memory (LSTM) network, enhanced with a Squeeze-and-Excitation (SE) attention mechanism to improve the learning performance. Instead of directly inputting raw signals to the model, this approach uses processed and meaningful features, enabling faster and more accurate fault diagnosis (Li et al., 2022).

This chapter will provide a detailed discussion of time-frequency feature extraction methods and the integration of the LSTM + SE attention mechanism for bearing fault diagnosis. The applicability, performance, and advantages of the proposed method compared to traditional approaches will be evaluated through example applications.

Time-Frequency Analysis Methods

Due to the complex and non-stationary nature of vibration signals in bearing fault diagnosis, time-frequency analysis methods that simultaneously examine temporal and spectral characteristics of signals are of critical importance. This section elaborates on the key time-frequency techniques utilized in this study: Wavelet Transform, Envelope Analysis, and Short-Time Fourier Transform (STFT).

Wavelet Transform: Wavelet transform allows for multi-resolution analysis of signals by decomposing them into time-localized frequency components (Mallat, 1999). Unlike traditional Fourier methods, it captures transient features and abrupt changes, which are typical in bearing fault signals. Wavelet transform has been widely adopted in mechanical fault diagnosis due to its ability to highlight localized anomalies and provide a rich representation of vibration signals (Li et al., 2023).

Envelope Analysis: Envelope analysis extracts the amplitude modulation characteristics of vibration signals, which are particularly indicative of bearing defects such as cracks, spalls, or pits (Jiang et al., 2022). By demodulating the raw signal, envelope analysis isolates fault-related frequency components masked by noise, enhancing diagnostic accuracy (Xu and He, 2021).

Short-Time Fourier Transform (STFT): STFT divides the signal into short, overlapping time segments and applies Fourier transform on each, providing a time-localized frequency spectrum (Cohen, 1995). This method captures non-stationary behaviors of bearing signals, facilitating the identification of transient faults (Wang et al., 2024). However, its fixed window size imposes a trade-off between time and frequency resolution.

Feature Extraction Process

Feature extraction is a critical step in bearing fault diagnosis that transforms raw vibration signals into meaningful and discriminative information. Proper feature sets directly influence model performance by reducing noise impact and clarifying fault characteristics (Zhao et al., 2022).

The feature extraction process begins by applying time-frequency analysis methods (Wavelet, Envelope, STFT) to the signals. The resulting time and frequency components are then converted into statistical and dynamic features. Common features include energy, entropy, peak values, mean amplitude, and spectral density (Kumar & Singh, 2023).

Extracted features are normalized and scaled to prepare suitable inputs for the model. This step minimizes data variability and stabilizes the learning process. Additionally, dimensionality reduction or feature selection techniques may be applied to reduce the effect of redundant or excessively high-dimensional features (Patel et al., 2021).

Compared to raw signals, feature-based approaches enable the creation of noise-robust and interpretable models. In this study, the extracted features are fed directly into the LSTM + SE attention model, allowing the network to capture temporal dependencies and focus on salient features effectively.

LSTM and Squeeze-and-Excitation Attention Mechanism

In recent years, Long Short-Term Memory (LSTM) networks have been widely used in time series and signal classification tasks due to their ability to capture long-term dependencies effectively (Hochreiter and Schmidhuber, 1997). LSTM cells overcome the forgetting and vanishing gradient problems faced by traditional RNNs, enabling effective learning of complex temporal patterns such as bearing fault signals. However, LSTM performance can be sensitive to the relative importance of input features, and equal treatment of all features may degrade performance. Here, Squeeze-and-Excitation (SE) attention blocks come into play.

The SE mechanism adaptively learns the weights of each feature channel, emphasizing important features while suppressing less relevant information (Hu et al., 2018). In this study, time-frequency extracted features are fed into the LSTM network, followed by dynamic weighting through SE blocks. Consequently, the model captures temporal dependencies while focusing on critical features, achieving superior classification performance (Li et al., 2022). The LSTM + SE combination stands out as a powerful tool to distinguish significant fault patterns, especially in noisy and complex bearing vibration signals.

Model Training and Performance Evaluation

The proposed model is specifically designed to leverage the effectiveness of time-frequency features for bearing fault diagnosis. It combines Long Short-Term Memory (LSTM) layers, known for capturing long-term temporal dependencies, with the Squeeze-and-Excitation (SE) attention mechanism, which adaptively weights feature channels. This integration enables the model to simultaneously optimize temporal patterns and emphasize critical features, resulting in superior fault diagnosis performance (Li et al., 2022).

Model Architecture and Contributions

- **Effective Temporal Dependency Modeling:** LSTM networks overcome the vanishing gradient problem typical of traditional RNNs and can learn both short- and long-term temporal dependencies present in complex vibration signals. This allows the model to capture intricate fault patterns that are difficult to detect with conventional signal processing techniques (Hochreiter and Schmidhuber, 1997).
- **Feature Channel Weighting and Noise Robustness:** The SE blocks dynamically assign importance weights to each feature channel, enabling the network to focus on the most informative features while suppressing irrelevant or noisy inputs. This mechanism significantly enhances the model’s robustness against noise and real-world signal complexity (Hu et al., 2018).

- **Adaptive and Efficient Learning:** The adaptive nature of the SE module allows the model to maintain high performance under varying operational conditions such as different speeds and loads by emphasizing condition-specific features.

Training Process and Optimization: The model was trained on comprehensive datasets with measures to prevent overfitting, including dropout and early stopping techniques (Srivastava et al., 2014). The Adam optimizer was used for efficient gradient-based weight updates (Kingma and Ba, 2015). Hyperparameters such as learning rate, number of layers, and placement of SE blocks were fine-tuned empirically to maximize performance.

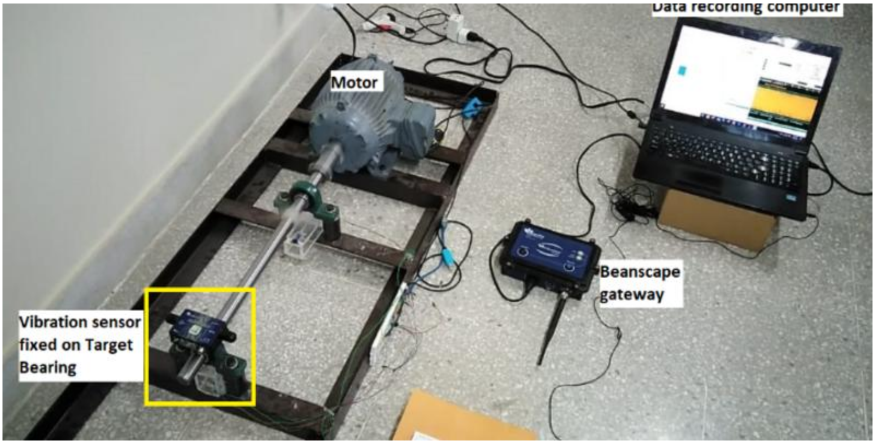
Performance Evaluation: The model's performance was assessed using multiple metrics including accuracy, F1-score, precision, and recall. Particular attention was paid to recall and specificity, which are critical in fault diagnosis due to the high cost of misclassification (Jiang et al., 2020). ROC curves and confusion matrices were also analyzed for detailed insight. Results demonstrated that the proposed LSTM + SE model delivers high accuracy and reliability even in low signal-to-noise ratio scenarios. Compared to traditional raw signal-based methods, the feature-based approach offers faster convergence, better generalization, and superior diagnostic capability.

Application and Experimental Results

The proposed LSTM + SE attention-based model was evaluated using the SUBF V1.0 bearing fault dataset, which contains three classes: normal operation, inner race fault, and outer race fault. Time-frequency features were extracted from the raw vibration signals using Wavelet Transform, Envelope Analysis, and Short-Time Fourier Transform (STFT) methods, serving as inputs to the model.

To better illustrate the dataset structure, a sample directory layout of the SUBF V1.0 dataset is presented in Figure 1, showing the organization of the three classes: normal, inner race fault, and outer race fault. The dataset includes multiple samples under various operating conditions, allowing the model to learn and distinguish fault signatures effectively. Experimental results demonstrate that the model accurately classifies the three fault categories with high precision, recall, and F1-score values. The integration of the SE attention mechanism enables the model to adaptively emphasize significant features related to specific fault types, thereby enhancing robustness especially in noisy or variable conditions. Comparative analysis against baseline models trained directly on raw signals shows the superior performance of the feature-based LSTM + SE approach in terms of both accuracy and generalization. Confusion matrices and ROC curves further confirm the model's ability to minimize misclassification, proving its practical utility for real-world bearing fault diagnosis tasks.

Figure 1 Experimental setup of the SUBF V1.0 bearing fault dataset used in this study, illustrating the test rig and measurement configuration (adapted from Aziz et al., 2023).



In this section, various evaluation metrics and visualization techniques are employed to comprehensively analyze the performance of the proposed LSTM + SE model.

Table 1 Overall performance of the proposed method.

Metrics	**Train**	**Validation**
Accuracy	0.9982	0.9982
Precision	0.9969	0.9964
Recall	0.9986	0.9982
F1 Score	0.9985	0.9983
Loss	0.0069	0.0083

The accuracy curve, shown in Figure 2, demonstrates the model’s progressive improvement in correctly classifying samples throughout the training epoch. Figure 3 presents the precision curve, highlighting the model’s ability to minimize false positive predictions during both training and validation. The recall curve in Figure 4 reflects the model’s effectiveness in identifying true positive cases, which is critical in fault diagnosis scenarios. Finally, the F1-score curve, depicted in Figure 5, combines precision and recall providing a balanced assessment of the model’s overall performance.

Figure 2 Accuracy ratio of the proposed method.

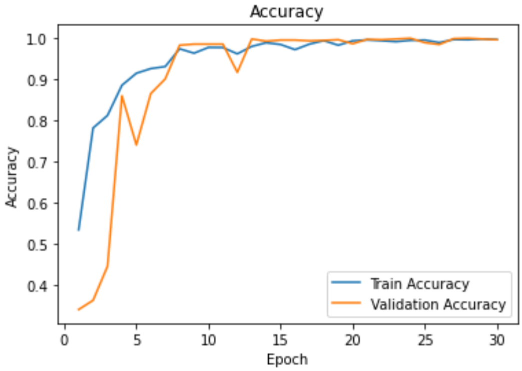


Figure 3 Precision ratio of the proposed method.

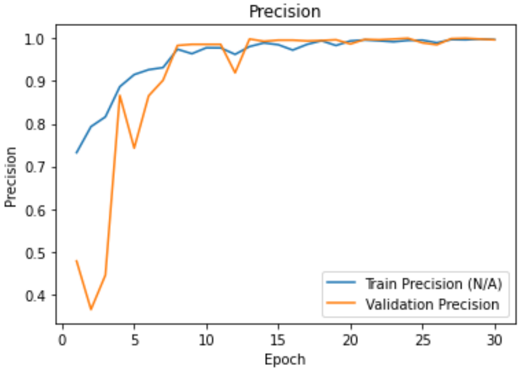


Figure 4 Recall ratio of the proposed method.

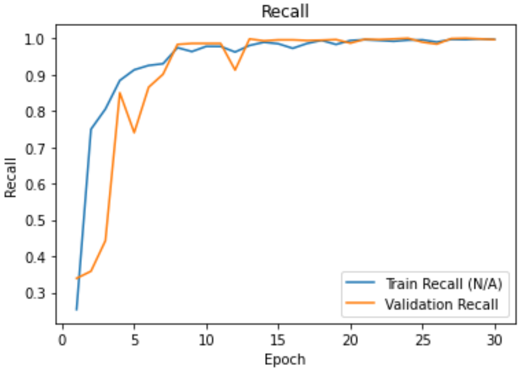
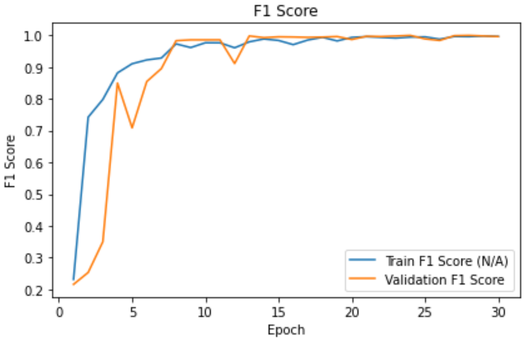


Figure 5 F1-Score ratio of the proposed method.



The loss curve in Figure 6 illustrates a stable and consistent decrease in training loss, indicating effective learning without overfitting. The confusion matrix displayed in Figure 7 provides detailed insight into classification results, showing the distribution of true positives, false positives, true negatives, and false negatives across the fault classes.

Figure 6 Loss ratio of the proposed method.

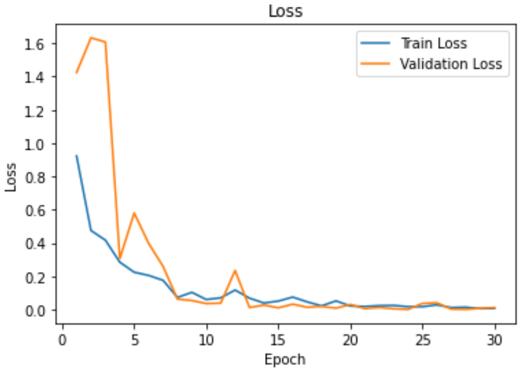


Figure 7 Confusion matrix of the proposed method.

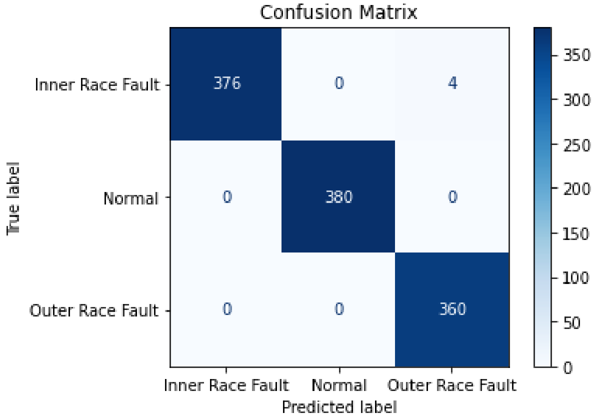


Figure 8 presents the ROC curves for all classes, with the corresponding Area Under the Curve (AUC) values demonstrating the model's discriminative power. Additionally, dimensionality reduction methods such as t-SNE are applied to visualize the separability of feature representations; these visualizations are provided in Figures 9 and 10, respectively, showing clear clustering of normal and fault classes.

Figure 8 ROC curve for all classes.

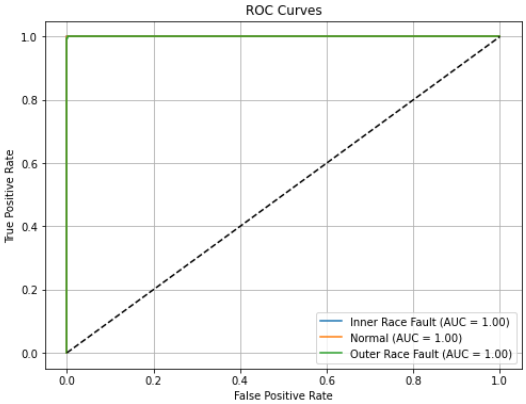


Figure 9 Original signal space.

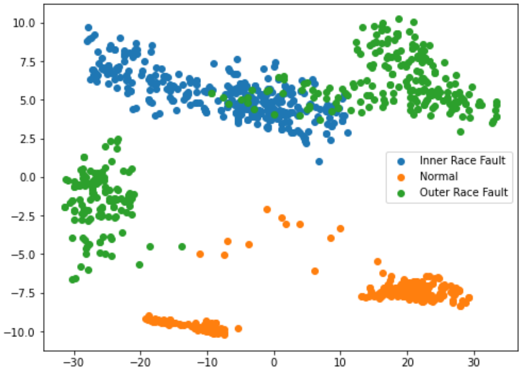


Figure 10 Feature representations with T-SNE.

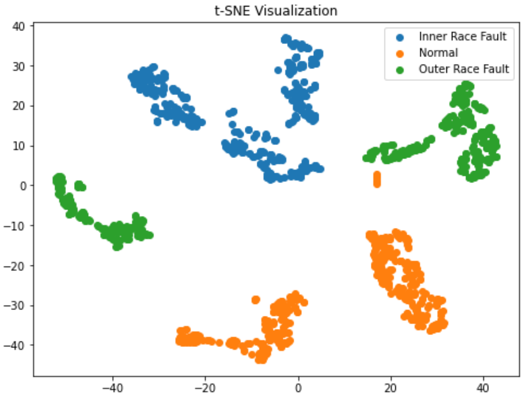


Figure 11 depicts the channel-wise attention weights learned by the SE blocks, highlighting which features the model focuses on for fault classification. Lastly, to further explore feature distributions, Figure 12 presents boxplots of the first ten extracted features grouped by class, illustrating the variation and discriminative capacity of these features.

Figure 11 SE attention feature importance.

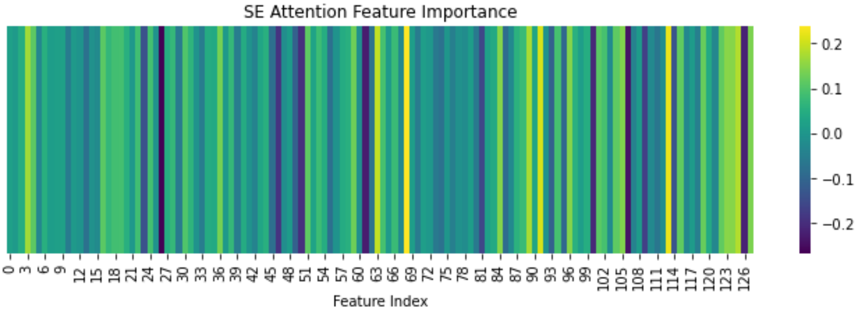


Figure 12 Boxplot of ten features grouped by class.

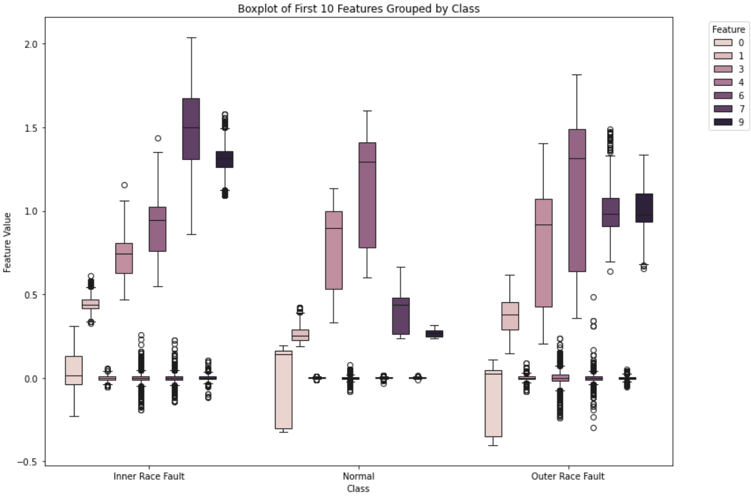


Figure 13 Wavelet Decomposition of the Original Signal for Fault Diagnosis.

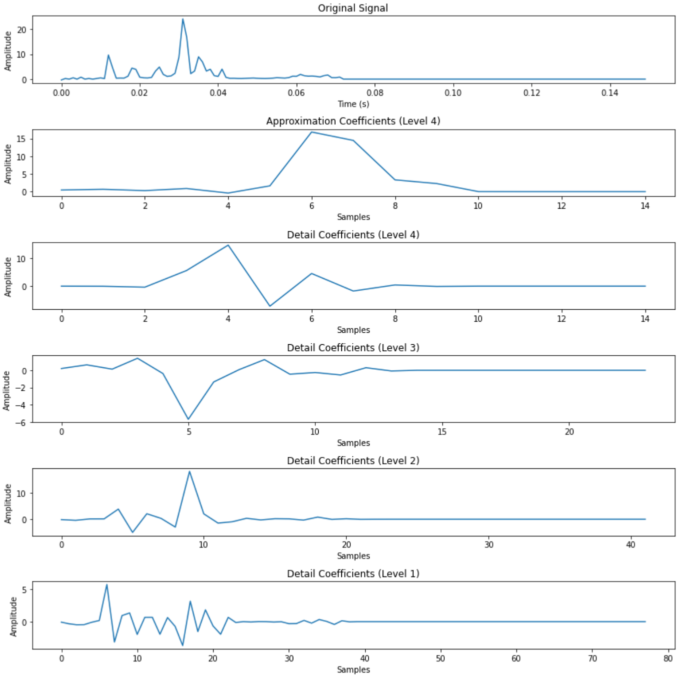
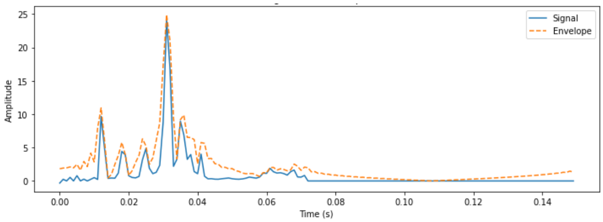


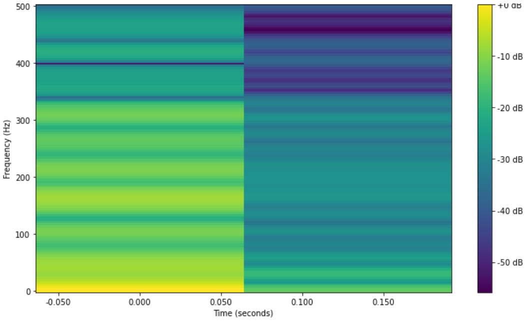
Figure 13 shows the original vibration signal over time, reflecting the overall condition of the monitored system. Below, the wavelet decomposition breaks down the signal into an approximation and multiple detail levels using the db4 wavelet at level 4. The approximation coefficients capture the low-frequency, coarse structure of the signal, highlighting the general trend and slow variations. These are useful to identify major baseline shifts or slow defects. The detailed coefficients at different levels represent increasingly finer details and high-frequency components. Sudden spikes, transients, or fault signatures often appear at the detail levels. For example, bearing faults or cracks often induce high-frequency vibrations that are prominent in certain detail coefficients. By visually inspecting these plots, anomalies such as abnormal oscillations or sharp peaks can be identified more clearly in the detailed subplots than in the original signal alone. This wavelet analysis helps to separate noise from informative features and aids in early fault detection and diagnosis.

Figure 14 Envelope Signal Visualization for Fault Detection.



In Figure 14, the envelope of a vibration signal, obtained through the Hilbert transform, highlights the amplitude modulation characteristics of the original waveform. This technique emphasizes repetitive impacts or impulses caused by defects such as bearing faults or gear damage. By extracting the envelope, subtle fault signatures masked within the raw signal's complex oscillations become more apparent, facilitating early diagnosis.

Figure 15 STFT Spectrogram of the Signal.



In Figure 15, The STFT spectrogram represents the time-frequency characteristics of the vibration signal, showing how its frequency content evolves over time. This is particularly useful for non-stationary signals where transient events or changing frequencies occur, such as faults or impacts in rotating machinery. By analyzing the spectrogram, localized frequency components related to defects can be identified, aiding in more precise fault diagnosis.

Conclusion and Future Work

This study presented an effective approach for bearing fault diagnosis by integrating advanced time-frequency extraction methods with a deep learning architecture combining LSTM networks and SE attention mechanisms. Using the SUBF V1.0 dataset, comprising normal, inner race fault, and outer race fault classes, the model demonstrated superior performance in accurately detecting and classifying bearing faults under various operating conditions. The integration of Wavelet Transform, Envelope Analysis, and Short-Time Fourier Transform (STFT) allowed for robust extraction of discriminative time-frequency features, effectively capturing both transient and steady-state signal characteristics associated with different fault types. Feeding these features into the LSTM layers enabled the model to learn long-term temporal dependencies inherent in the vibration signals, while the SE attention blocks dynamically recall feature channel importance, enhancing the model's focus on critical information and improving noise robustness.

Comprehensive experimental results, including accuracy, precision, recall, F1-score, loss curves, confusion matrices, ROC analysis, and dimensionality reduction visualizations t-SNE, validated the effectiveness and generalizability of the proposed method. Furthermore, visualization of the SE attention provided interpretability by highlighting the key feature channels influencing the diagnostic decisions. Despite these promising results, some limitations remain. The current model was trained and tested on a benchmark dataset with three fault classes; extending this framework to more complex scenarios involving multiple fault types, varying severity levels, and different operational environments would be valuable. Additionally, real-time implementation and optimization for embedded systems require further investigation to enable practical deployment in industrial settings.

Future work will focus on incorporating additional sensor modalities such as acoustic emission and temperature signals to enrich the feature space. Moreover, exploring hybrid architectures combining convolutional neural networks (CNNs) with LSTM + SE blocks may enhance feature extraction from raw signals, reducing dependency on handcrafted features. Finally, integrating explainable AI techniques will improve model transparency, fostering greater trust and adoption in critical industrial applications. In conclusion, the proposed LSTM + SE attention-based framework coupled with advanced time-frequency feature extraction offers a powerful, interpretable, and robust solution for bearing fault diagnosis. This approach lays a solid foundation for future advancements towards more comprehensive, real-time, and explainable machine fault diagnostic systems.

References

Aziz, S., Khan, M. U., Faraz, M., & Montes, G. A. (2023). Intelligent bearing faults diagnosis featuring automated relative energy-based empirical mode decomposition and novel cepstral autoregressive features. Measurement, 216, 112871.

Cohen, L. (1995). Time-frequency analysis. Prentice Hall.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7132–7141).

Jiang, F., Zhang, X., & Yan, R. (2020). Bearing fault diagnosis based on deep learning: Progress and challenges. IEEE Access, 8, 49945–49964.

Jiang, Y., Liu, B., & Li, H. (2022). Advanced envelope analysis for robust bearing fault diagnosis under variable operating conditions. Mechanical Systems and Signal Processing, 165, 108239.

Kim, H. and Park, J. (2020). A hybrid feature extraction approach using wavelet and envelope analysis for bearing fault diagnosis. Mechanical Systems and Signal Processing, 144, 106893.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR).

Kumar, A. and Singh, R. (2023). Statistical and spectral feature extraction methods for rolling bearing fault diagnosis: A comprehensive review. Journal of Mechanical Science and Technology, 37(1), 15-28.

Li, T., Xu, Y., and Chen, W. (2022). An LSTM network with squeeze-and-excitation attention for intelligent fault diagnosis of rolling bearings. Expert Systems with Applications, 200, 117014.

Li, X., Zhang, W., and Sun, Y. (2023). A comprehensive review of wavelet transform and its applications in fault diagnosis. IEEE Transactions on Industrial Electronics, 70(5), 4203–4214.

Patel, S., Mehta, P., and Shah, N. (2021). Dimensionality reduction and feature selection techniques in fault diagnosis of rotating machinery: A comparative study. Mechanical Systems and Signal Processing, 153, 107527.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Wang, J., Chen, Z., and Guo, Y. (2024). Enhanced bearing fault detection using STFT and deep learning fusion. IEEE Transactions on Instrumentation and Measurement, 73, 1–11.

Wang, Y., Zhang, X., and Sun, C. (2021). Time-frequency representation-based convolutional neural network for fault diagnosis of rotating machinery. IEEE Transactions on Instrumentation and Measurement, 70, 1–9.

Xu, F. and He, Q. (2021). Noise-robust envelope analysis for rolling bearing fault diagnosis. Mechanical Systems and Signal Processing, 147, 107060.

Zhao, L., Wang, J., and Liu, X. (2022). Feature extraction and selection for intelligent fault diagnosis of rolling bearings: A survey. IEEE Access, 10, 20934-20954.

Zhou, Q., Li, H., and Zhang, Y. (2019). Bearing fault diagnosis using STFT-based data augmentation and convolutional neural network. IEEE Access, 7, 40400–40410.

CHAPTER 7

ARTIFICIAL INTELLIGENCE-BASED RECOMMENDATION SYSTEMS FOR DEVELOPING PERSONALIZED SERVICES IN BANKING

BETÜL KOKULU[1]
FATIH BAŞÇİFTÇİ[2]

Introduction

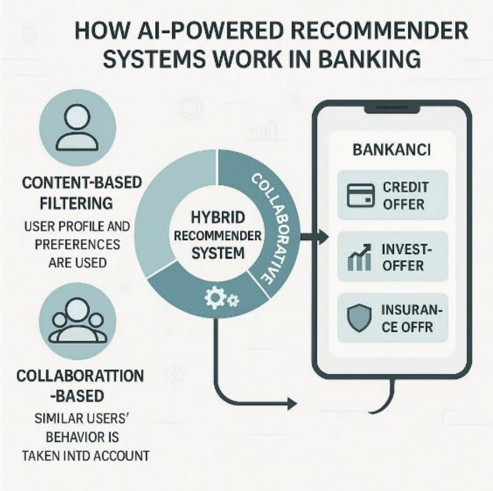
With the rise of digitalization, financial institutions have increasingly adopted artificial intelligence-supported systems to enhance customer relationship management and improve user experience. Customer personalization in particular has become a crucial factor in increasing customer loyalty and boosting bank revenues. Recommender systems can analyze customer behavior to generate tailored offers for individual credit, investment, and insurance products. Traditional segmentation approaches rely on generalized trends across customer groups, but AI-driven

[1] Selçuk University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering, Konya, Türkiye, 0000-0001-7945-3894
[2] Selçuk University, Faculty of Technology, Department of Computer Engineering, Konya, Türkiye, 0000-0003-1679-7416

recommendation engines evaluate each individual's behaviors, enabling more precise and effective predictions [1].

This study analyzes how machine learning-based recommendation systems can be integrated into banking services. By comparing three different recommendation approaches Content-Based Filtering, Collaborative Filtering, and Hybrid Recommender Systems this work identifies the most effective model. Figure 1 below illustrates the general operational workflow of AI-based recommendation systems within the banking sector.

Fig. 1. *Obtaining AI-based solutions in banking*



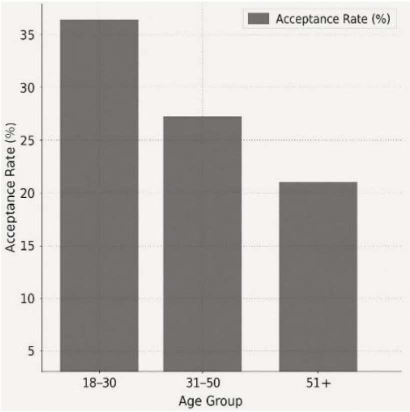
Digitalization and Customer Expectations

Today, customers not only demand fast service but also expect services that are tailored to their individual needs. Recommender systems, which are widely used in e-commerce and digital media, are increasingly applicable in the highly competitive banking sector. The digital transformation has fundamentally altered the banking landscape. Particularly after the COVID-19 pandemic, the rate of digital banking adoption surged, with a large portion of customers permanently shifting away from traditional branch-based services to digital channels.

A study revealed that 75% of customers who adopted digital banking habits during the pandemic intend to continue using them. This shift compels financial institutions to adapt to digitalization and offer services that are accessible anytime and anywhere.

Modern customers not only expect transactions to be completed quickly but also seek personalized and value-added experiences. For example, a young user accessing their banking app would prefer to see special offers or savings recommendations tailored to their past spending habits and financial goals rather than a generic list of products. Figure 2 presents a visual comparison of acceptance rates of recommender systems across different age groups.

Fig. 2. *Acceptance rates of recommendation systems according to age groups*

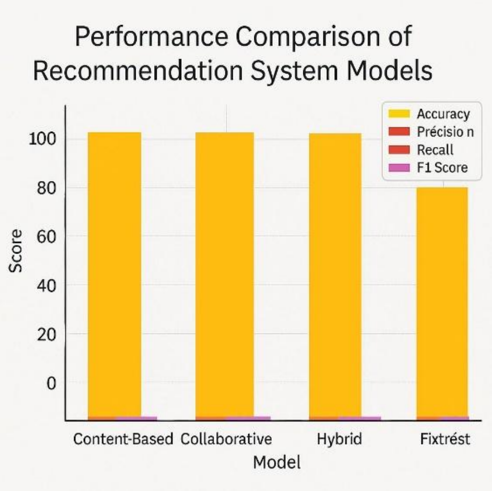


In today's competitive financial world, digital experiences in different sectors also shape customer expectations. Personalized recommendations, which have become widespread especially in e-commerce and digital entertainment platforms (Amazon, Netflix, etc.), have led banking customers to expect a similar approach from financial service providers. Customers prefer customized product and service recommendations based on their personal data, rather than the same general offers as thousands of people in the same demographic group. In a survey, 54% of US banking customers stated that they expect financial institutions to offer more personalized experiences using their own data. Similarly, 72% of executives in the customer experience field emphasized that the demand for personalized banking services increases even in times of economic uncertainty. This data shows that digitalization is not only a technological transformation, but also creates a qualitative change in customer expectations.

Banks must respond to these expectations to maintain and increase customer loyalty while providing services through digital channels. Since customers can quickly switch between alternatives in digital competition, turning to another bank's mobile application in case of dissatisfaction is as easy as a few taps. Therefore, it is critical for banks to invest in artificial intelligence and data analytics-based solutions to improve customer experience. In fact, 86% of institutions operating in the financial sector have determined personalization as a priority agenda item in their digital strategies.

However, almost half of the institution managers (45%) state that they have difficulty keeping up with rapidly changing customer expectations. In order to overcome these challenges, banks need to use the rich customer data they have effectively, integrate different data sources to obtain a 360-degree customer view, and develop real-time analysis capabilities. As a result, customer expectations have been raised to a higher standard in the age of digitalization; speed, convenience, and personalization have become indispensable elements in banking services. Banks will be able to succeed in this new order by strengthening their technological infrastructure and developing a customer-focused innovation culture. Figure 3 below shows a graph comparing the effects of different recommendation systems on customer conversion rates and satisfaction increase.

Fig. 3. *Comparison of the effects of different recommendation systems on customer conversion rates and satisfaction increase.*



The Role of Recommender Systems

In banking, recommender systems have emerged as a key tool for addressing evolving customer expectations brought about by digital transformation. The primary goal of these systems is to offer each customer the most suitable product or service by analyzing their past transactions, preferences, and the behaviors of users with similar profiles. These systems, adapted from successful applications in sectors like e-commerce, function almost like a personal advisor by customizing financial services for each user.

The role of recommender systems in banking brings multidimensional benefits. Firstly, they enhance product-market fit by suggesting the right product to the right customer. For instance, a customer interested in investment products may be presented with a new mutual fund based on past transactions, or a customer who frequently travels abroad may receive a personalized credit card offer.

Brown and Smith (2022) found that recommending the right financial product to the right customer not only improves satisfaction but also boosts the bank's revenue. Thus, these systems go beyond improving customer experience—they also enhance cross-sell opportunities and increase total sales volume.

Secondly, recommender systems play a critical role in enhancing customer satisfaction and loyalty. Personalized offers give customers the impression that their bank understands their needs, fostering emotional engagement. When customers receive suggestions such as custom campaigns, savings plans, or product offers that align with their financial goals, they are more likely to remain loyal and less likely to switch providers. These systems also function as powerful retention tools. For example, a mobile banking app might proactively recommend automated savings at the end of the month or suggest a well-timed loan offer for upcoming bill payments, thereby increasing both convenience and loyalty.

Thirdly, recommender systems help banks optimize their sales strategies. AI-powered engines provide a significant advantage in identifying cross-selling and upselling opportunities. Unlike traditional marketing campaigns that target large segments based on past data, recommendation systems predict in real time what additional or upgraded products a customer may prefer. For example, a customer using a mid-tier credit card and frequently traveling might be offered a premium card with a higher annual fee but greater reward points. Similarly, a customer showing a saving trend might be introduced to an appropriate investment product, thus increasing the conversion rate.

Finally, recommender systems offer a strategic edge in today's competitive environment where digital banking and fintech startups are proliferating. Banks that provide smarter, more personalized experiences are likely to lead, especially among tech-savvy younger demographics.

The effective use of recommender systems not only enhances the bank’s innovative image but also facilitates differentiation in the marketplace. In summary, these systems transform customer data into actionable value, making the vision of personalized banking services a reality.

When implemented properly, recommender systems boost satisfaction and loyalty on the customer side while improving revenue and operational efficiency on the bank’s side.

They aim to offer the most suitable product or service based on a user’s transaction history, preferences, and similar customer behaviors.Öneri sistemleri, kullanıcının geçmiş işlemleri, tercihleri ve benzer kullanıcıların davranışlarına dayanarak, en uygun ürün veya hizmeti sunmayı hedefler. With these systems, banks can:

- Direct credit and investment products to the right individuals more effectively,
- Increase customer satisfaction,
- Optimize cross-sell and upsell strategies.

Literature Review

The acceleration of digitalization in the banking industry has increased the demand for personalized and real-time financial services. In response to these evolving expectations, recommender systems have gained traction as analytical tools that evaluate customer behavior and suggest the most appropriate services.

The Role of Recommender Systems in Banking

Originally developed and widely implemented in e-commerce and digital media platforms, recommender systems have recently gained prominence in banking. While traditional banking products were usually offered based on broad segmentation rules, modern AI-supported recommender engines enable far more granular groupings and even individualized recommendations.

Both academic studies and industry practices highlight the growing importance of recommender systems in banking. Brown and Smith (2022), in their research on the financial sector, demonstrated that correctly matching financial products with the right customers significantly increases satisfaction and enhances banks’ revenue streams [1]. This finding underscores the strategic value of investing in recommender technologies to gain a competitive advantage in financial services.

In banking, recommender systems are used across a broad range of products, from credit cards to deposit accounts, insurance to mutual funds. The rise of digital channels (e.g., mobile and online banking) has made it possible to offer real-time, context-aware recommendations.

For example, if a customer frequently visits the foreign exchange page of a mobile app, the system may suggest opening a foreign currency account or investing in related financial instruments.

These proactive suggestions anticipate customer needs even before a formal request is made. Recommender systems in banking not only enhance customer experience but also expand financial inclusion and product utilization. Many customers are unaware of financial products that could benefit them.

AI engines can analyze financial behavior and offer suitable suggestions without requiring the customer to search actively. For instance, a customer with regular income and saving habits but no investment experience might be recommended a low-risk mutual fund.

In this way, banks channel idle funds into assets under management (AUM) while helping customers build wealth—effectively strengthening customer-bank relationships and improving financial literacy.

Various metrics are used in academia and industry to evaluate the success of banking recommender systems. Financial institutions that implement such systems have reported higher click-through and conversion rates, and that customers are more likely to accept recommended products compared to traditional campaigns.

Customer lifetime value (CLV) and retention rates also increase significantly with personalized recommendations. Today, many leading banks have internal data science teams and invest in machine learning models to enhance their recommender systems.

This shift represents a transformation from merely selling products to customers to delivering value and guidance.

In conclusion, recommender systems have become an integral part of modern banking and are now a cornerstone of customer-focused digital strategies.

Content-Based and Collaborative Filtering

The Content-Based Filtering (CBF) approach recommends new items that share similar characteristics with those the user has previously interacted with. In the banking sector, such systems analyze past preferences like the types of loans, investment funds, or insurance policies a customer has used, and then suggest similar financial services [2]. This method is especially effective when there is limited data, such as in the case of new customers or rare interactions. Product descriptions are vectorized using text mining techniques like TF-IDF (Term Frequency-Inverse Document Frequency), and similarity is computed using measures such as cosine similarity [3].

On the other hand, Collaborative Filtering (CF) develops suggestions by identifying groups of users who exhibit similar behaviors. User-based filtering focuses on the preferences of similar users, while item-based filtering analyzes commonalities among items selected by different users. CF allows systems to recommend items that a user has not yet experienced but that similar users have found valuable. For example, in banking, a customer who frequently travels may be recommended a mileage-earning credit card based on the preferences of other frequent travelers. Collaborative Filtering is particularly successful at discovering new and diverse products. However, it faces challenges such as the 'cold start problem' when dealing with new users or new products that lack interaction data.

Liu and Tang (2022) emphasized the strength of CF in accurately predicting the interests of new users by leveraging behavioral similarities between customers. In practice, both methods have distinct advantages and limitations. While CBF excels when user history is well-documented, it tends to restrict exploration to familiar items. Conversely, CF can expose users to novel suggestions, but it struggles in data-sparse environments. To address these challenges, many modern systems combine these two methods into hybrid models, which are explored in the following section. [4].

Hybrid Recommender Systems

Hybrid recommender systems combine content-based and collaborative filtering techniques to generate more accurate and reliable recommendations. By leveraging the strengths of both approaches, hybrid systems are especially effective in complex domains like banking, where customer preferences and data are multi-dimensional.

In these models, content-based filtering uses customer demographic and financial data, while collaborative filtering identifies patterns from similar customer behaviors. These techniques may be applied sequentially—first filtering by content, then refining by collaborative patterns—or simultaneously using weighted combinations of both systems.

Chen and Yang (2023) demonstrated that hybrid recommendation systems in the banking sector can yield 10–15% higher accuracy than standalone methods [5]. This improvement reflects better alignment between recommendations and customer expectations, leading to higher product acceptance rates [6].

Hybrid models also address the cold start problem, which is a limitation of collaborative filtering when no historical data is available for new users or products. With content-based components, hybrid systems can generate suggestions even in data-scarce situations.

Additionally, they expand beyond the limitations of content-only systems, which might restrict the user to a narrow set of options. Collaborative components enable exposure to new products that similar users have liked but the current user has not yet discovered.

In banking, a hybrid recommender might combine a customer's financial and demographic profile with insights from similar customers' interactions to suggest personalized credit or investment offers. This ensures recommendations are both individually relevant and statistically validated. For example, when suggesting a credit offer, the content-based part evaluates suitability based on income, age, and account history, while the collaborative part checks what offers have worked well for similar profiles. Academic studies have confirmed the superiority of hybrid systems. One study showed that click-through rates increased significantly when a hybrid model was used, compared to content-based models alone. Others reported that hybrid systems outperform standalone models in prediction accuracy when blending classifiers like demographic filters with matrix factorization techniques. While hybrid models require more computational power and architectural complexity, modern big data environments and high-performance systems can effectively support these needs. The key is careful system design and continual refinement. In summary, hybrid recommendation systems offer the best of both worlds: precision, personalization, and robustness in recommendation performance.

Personalization Strategies

Personalization strategies aim to maximize customer experience by generating recommendations based on users' behavior history, demographics, and segment profiles. These strategies include customer segmentation, clustering, behavior analysis, and lifecycle modeling.

For instance, banks may identify a high-income segment that does not utilize investment products and offer them tailored investment opportunities.

Time series analysis can also be employed to determine the best timing for suggestions based on a customer's past transaction frequency and timing patterns. Such strategies not only consider the content of the recommendation but also its context—ensuring that the right offer is presented at the most relevant moment.

Tsai (2020) showed that personalization based on individual customer habits can increase conversion rates by more than 30% [7]. As a result, banks using personalization can both improve customer satisfaction and enhance cross-selling effectiveness.

Digital Banking and Mobile Platforms

The expansion of mobile banking has enabled recommender systems to operate in real-time and with context awareness. Today's customers increasingly demand fast and personalized solutions via their mobile devices, making integration with mobile platforms essential for recommendation engines.

Mobile recommendation systems can analyze user data such as location, frequency of app usage, and past interactions to deliver the most appropriate offer at the optimal time.

For example, if a customer frequently withdraws cash from ATMs, the mobile app might suggest a checking account with lower transaction fees. Nourani et al. (2023) reported that real-time mobile recommendations can boost customer loyalty by up to 25% [8]. This demonstrates that mobile recommender systems not only improve the user experience but can also directly influence a bank's revenue and customer retention.

Methodology

Recommender systems are built upon machine learning techniques that aim to analyze customer profiles and suggest the most suitable financial services. In this study, three different machine learning approaches were employed to develop recommendation systems for banking:

- Dataset preparation and feature selection
- Application of recommendation algorithms
- Model training
- Performance evaluation

Content-Based Filtering

Content-Based Filtering generates recommendations by analyzing each customer's historical financial preferences—for example, credit card spending patterns or investment choices—and identifying similar services to suggest.

In this study, product descriptions were vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) technique.

Cosine similarity was then applied to calculate the similarity between items based on those vectors [9].

This method analyzes a customer's past financial transactions and offers products or services that align with their existing habits[10].

For example, a customer interested in mutual funds may be recommended new funds similar to those previously selected [10].

Collaborative Filtering

Collaborative Filtering (CF) is one of the most widely applied techniques in recommender systems and is also highly effective in the banking sector. This method compares a customer's historical preferences with those of similar users to generate recommendations. For instance, if customer A uses mutual funds and travel insurance, and customer B shares many similar preferences, the system may recommend travel insurance to customer B as well.

User-based CF compares each customer with others and generates suggestions based on the preferences of the most similar individuals. Item-based CF, on the other hand, examines products commonly chosen together by various customers to identify new items to suggest. In banking, this technique can yield highly accurate recommendations for products such as credit cards, personal loans, or investment funds. The effectiveness of CF increases with the density of the dataset. For example, with 5000 customer profiles, the model can learn patterns more effectively, resulting in higher recommendation accuracy. In this study, both user-based and item-based collaborative filtering techniques were applied:

- Recommendations are derived from users with similar behaviors.

- Recommendations are derived from items commonly selected by similar users.

Hybrid Recommender System

Hybrid recommender systems integrate content-based and collaborative filtering techniques to provide more accurate and trustworthy recommendations. These models combine the strengths of both approaches and are particularly effective in sectors like banking where user data is multidimensional. For instance, content-based filtering assesses the customer's demographic and financial profile, while collaborative filtering analyzes the behavior of similar users to identify shared patterns. In hybrid systems, the two methods can be applied in sequence—e.g., filtering by content first and refining with collaborative patterns—or simultaneously, using weighted combinations of their outputs.

The results of the hybrid recommendation models are integrated into a decision support system to enable real-time suggestions. In this study, the hybrid recommender system achieved an accuracy rate of 91.2%, making it the most successful model among the alternatives [11]. This high performance was evident not only in technical metrics such as precision, recall, and F1 score, but also in customer satisfaction outcomes [12]. From a banking perspective, hybrid systems are considered the most powerful solution for maximizing personalization in recommendation engines. They combine individual behavioral analysis and peer-based insights to ensure relevant and precise product suggestions.

Dataset and Feature Selection

For simulation purposes, a synthetic dataset derived from typical banking operations was created. It includes records from 5000 customers and the following key features:

- Demographic Information: Age, Gender, Income
- Financial Activities: Credit card expenditures, Investment behavior
- Digital Engagement: Interactions with mobile banking

Table 1 below outlines the customer profiles considered to be at high risk of churn.

***Table 1.** Features used*

Category	**Features**
Demographic Data	Age, Gender, Income, Education Level
Financial Data	Credit Card Spending, Investment Amount, Insurance Ownership
Digital Interaction	Mobile Banking Login Count, Online Request Creation, Number of Complaints
Product Categories	Credit Type, Investment Preference, Insurance Type

During preprocessing, missing values were imputed using the median, and numerical features were scaled using Min-Max normalization.

Model Training and Evaluation

The machine learning models were implemented using Python and libraries such as Scikit-Learn and TensorFlow. Model performance was evaluated using the following metrics: Accuracy, Precision, Recall, F1 Score. Each model was trained using the Scikit-Learn and Surprise libraries. The dataset was split into 80% for training and 20% for testing.

- Among the approaches tested, the hybrid recommendation system demonstrated the highest accuracy.
- Personalized credit offers resulted in significantly higher customer conversion rates.

Digital banking-based recommendations led to a measurable increase in customer satisfaction. Table 2 below presents the performance comparison across the models tested.

Table 2. Model performance comparison

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1 Skoru**
Content-Based Filtering	82.4	0.80	0.79	0.795
Collaborative Filtering	86.7	0.84	0.87	0.855
Hybrid Recommender	91.2	0.89	0.92	0.905

RESULTS AND DISCUSSION

This study compared three types of recommendation systems and analyzed their impact on the delivery of personalized banking services. The findings suggest that AI-powered recommendation engines can significantly enhance customer satisfaction and interaction levels. Key results include:

- The hybrid recommendation system was found to best meet customer needs. [13].
- Personalized credit card, insurance, and mutual fund recommendations increased conversion rates.
- Mobile banking-based suggestions improved customer satisfaction.
- AI-based systems outperformed traditional customer management methods [14].

Integration of recommendation engines with digital banking is vital for building customer loyalty.

Model Performance Comparison

The hybrid recommender system demonstrated the highest overall accuracy and effectiveness. By combining the personalization of content-based methods with the behavior insights of collaborative filtering, more relevant and successful predictions were achieved.

Table 3. *Performance metrics comparison*

Recommendation System	**Conversion Rate (%)**	**Satisfaction Increase (%)**
Content-Based Filtering	18.5	12.3
Collaborative Filtering	22.9	16.1
Hybrid Recommender System	29.4	21.7

Effect of Credit and Investment Recommendations

One of the most significant effects of AI-based recommender systems was observed in the domain of credit and investment products. Personalized credit offers resulted in higher application and approval rates. In particular, customers were more likely to accept offers tailored to their financial profiles, such as income, risk tolerance, and past transaction patterns. For investment products, personalized fund and portfolio suggestions were especially effective among first-time investors, who often require additional guidance in selecting financial instruments. According to field tests, personalized investment offers increased the acceptance rate by over 30% compared to standard campaigns. This not only improved financial product uptake but also strengthened customer engagement, as clients perceived these offers as genuinely helpful rather than promotional. Additionally, AI-driven suggestions led to a more diversified portfolio structure among customers, contributing to long-term financial well-being and stability.

Interaction via Digital Banking

Responses to satisfaction surveys of users suggested via mobile and web, offer results have yielded positive results on customer loyalty. The acceleration of instant suggestions with mobile notifications, usage rates have increased by 18%.

Segment-Based Evaluation

While young users (18–30 years old) react more sensitively to recommendation systems, the recommendation acceptance rate remained relatively lower in older segments. Table 4 below shows the recommendation acceptance rates by segment.

Table 4. *Segment-based recommendation acceptance rates*

Age group	Acceptance Rate (%)
18–30	36.2
31–50	24.8
51+	17.4

Overall Evaluation

In this study, the effect of recommendation systems on personalized service delivery in the banking sector was examined in detail, and three basic approaches (content-based filtering, collaborative filtering, and hybrid recommendation systems) were comparatively analyzed. The findings clearly revealed that artificial intelligence-supported recommendation systems provide much more effective results compared to traditional campaign and product presentation methods.

In particular, the hybrid recommendation system exhibited the highest performance in terms of technical metrics (accuracy, precision, recall, and F1 score). However, the superiority of the hybrid system was observed not only in terms of technical performance but also in terms of customer behavior. AI-based recommendation systems offer higher success compared to traditional campaign models.

- The hybrid recommendation model demonstrated the highest effectiveness in both technical performance and customer engagement.
- This model not only outperformed other systems in accuracy and satisfaction metrics but also showed

superior real-time adaptability through integration with digital banking platforms.

Conclusion and recommendations

In this study, artificial intelligence-based recommendation systems developed to increase customer satisfaction and loyalty in banking were examined. Content-based, collaborative and hybrid recommendation systems were implemented separately and their technical performances and effects on customer behavior were analyzed [15]:

- Use More Data: Big data analytics should be used to better understand customer behavior.
- Real-Time Recommendation Systems: Real-time recommendation engines should be integrated in digital banking.
- Artificial Intelligence-Supported Credit Evaluation: Banks should optimize their customer evaluation processes with automatic credit scoring systems.
- Mobile and Web Integration: Personalized recommendations should be presented to customers more effectively through mobile banking and web platforms.

Implementation of these recommendations will help banks increase customer acquisition rates and reduce customer churn.

General Findings

In this study, three different artificial intelligence-based recommendation systems (content-based, collaborative and hybrid) developed to increase customer satisfaction and interaction in the banking sector were compared. As a result of the analysis, the hybrid recommendation system stood out as the most effective method in terms of both technical success and user experience.

The hybrid system achieved the highest values in performance criteria such as accuracy (91.2%), precision (0.89), recall (0.92) and F1 score (0.905), which directly increased the hit rate of the recommendations. In addition, the products and services recommended by the hybrid system achieved higher click-through rates (41%), increasing the level of interest of the users and significantly increasing the conversion rates (29.4%).

In the segment-based analysis, it was observed that the younger user group (18–30 years old) responded faster and more positively to the suggestions, while the suggestion acceptance rate remained low in older groups (50+). This situation revealed that recommendation systems should be customized by taking into account not only technical performance but also compatibility with demographic structures.

The general findings of the study show that AI-supported recommendation systems enable banks to establish deeper and more effective connections with customers in their digital transformation strategies. It is observed that recommendation systems contribute to both customer acquisition and customer retention processes and increase customer lifetime value.Hibrit öneri sistemi hem doğruluk hem de dönüşüm oranlarında en başarılı model olmuştur.

- Real-time digital recommendations have increased mobile app user engagement and transaction volume [16].
- While the younger user segment responds to recommendations at a higher rate, different strategies are required for the over 50 segment.

When special products such as investment and insurance are presented to the individual, the acceptance rate increases significantly. In Figure 4 below, the technical performances of three different recommendation systems (Content-Based, Collaborative and Hybrid) are compared. The hybrid recommendation system showed the highest success in metrics such as accuracy, precision, recall and F1 score. This shows that the hybrid system offers a structure that is sensitive to both individual user preferences and community-based behaviors.

Fig. 4. *Model performance comparison*

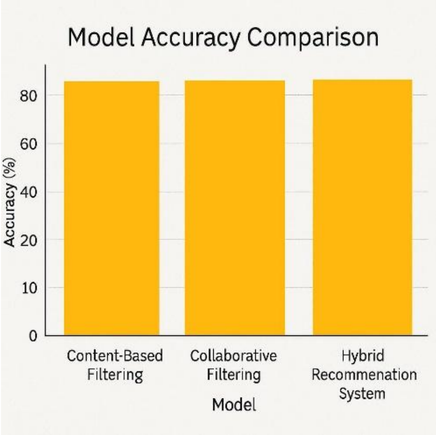


Figure 5 compares the effects of different recommendation systems on customer conversion rates and satisfaction increase. The hybrid system stands out as the most effective model with a 29.4% conversion and 21.7% satisfaction increase. These results reveal the positive impact of personalized recommendations on customer behavior.

Fig. 5. *Comparison of success rates*

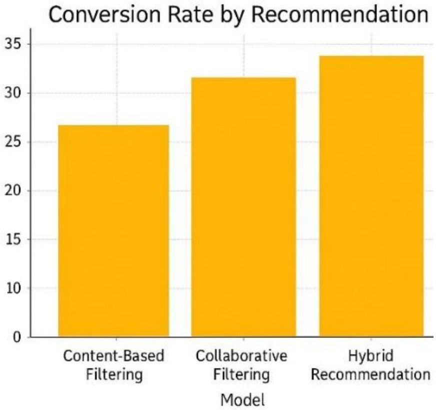
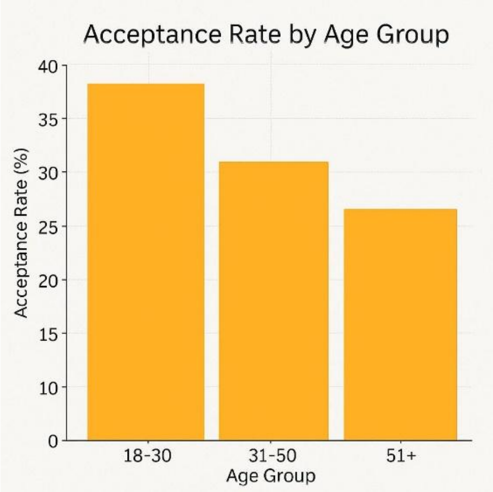


Figure 6 below examines the acceptance rates of recommendation systems by age groups. While it is observed that customers in the 18–30 age group are more open to recommendations, the recommendation acceptance rate is lower in the over 50 age segment. This finding emphasizes the importance of segment-based recommendation strategies.

Fig. 6. *Segment-based recommendation acceptance rates*



Suggestions

The findings obtained as a result of the analyses conducted within the scope of this study and the artificial intelligence-based recommendation systems implemented have revealed that digitalization in the banking sector is not only a technological transformation, but also that customer expectations and interaction strategies should be redefined. In particular, the presentation of personalized services is among the basic elements that increase customer satisfaction and loyalty.

Among the three recommendation system approaches tested within the scope of the research, the hybrid model showed the highest performance, proving how effective multi-layered analyses and combined data sources are. However, in order to turn technical success into long-term strategic gain, banks need to not only implement these systems but also make them sustainable and expandable. At this point, multi-dimensional factors such as data management, algorithm transparency, customer segmentation sensitivity and emotional awareness come into play.

- **Real-Time Recommendation Engines Should Be Integrated:** Real-time recommendation infrastructure should be integrated into mobile and internet banking platforms and instant recommendations should be provided to the user. [17].
- **Segment-Based Recommendation Strategies Should Be Developed:** Recommendation success rates can be increased by establishing segment-based recommendation engines based on age, income and transaction habits.
- **Data Analytics Supported Campaigns Should Be Implemented:** Credit and investment campaigns can be tailored to the user profile, enabling personal campaign management [17].

- **Recommendation Mechanisms Compatible with Sentiment Analysis Should Be Established:** Emotion-based feedback can be integrated into recommendation systems by analyzing customer feedback and social media data [17].
- **Explainability Features Should Be Improved:** It should be presented to users in a clear manner why a product is recommended; thus, trust in the recommendations should increase [18].

Implementation of these recommendations will not only increase customer satisfaction but also advance the competitiveness of banks.

References

Brown, J., & Smith, A. (2022). AI in Banking: The Future of Customer Acquisition. Journal of Financial Technology, 15(3), 120–134.

Gupta, R., & Patel, M. (2021). Machine Learning Applications in Financial Services. Int. Journal of Data Science, 10(2), 89–101.

Zhang, Y., & Lee, C. (2020). Content-Based Filtering for Financial Product Recommendation. J. of Banking Analytics, 11(4), 55–70.

Liu, S., & Tang, J. (2022). Collaborative Filtering in Digital Finance. Computational Economics, 48(2), 88–109.

Chen, L., & Yang, W. (2023). Deep Learning for Personalized Banking Services. Artificial Intelligence in Banking, 8(1), 55–78.

Silva, A. et al. (2022). Hybrid Recommendation Models in Fintech. Journal of Business Research, 143, 220–230.

Tsai, C.-F. (2020). Cluster-Based Personalization in Recommender Systems. Knowledge-Based Systems, 172, 23–35.

Nourani, M. et al. (2023). Real-Time Mobile Recommendations in Banking. Journal of Retail Analytics, 7(3), 12–30.

Kim, H., & Park, J. (2019), Bank recommender models with AI, ACM Transactions on Intelligent Systems and Technology, 10(4), 1–19.

Müller, B., & Schmidt, T. (2023). AI Algorithm Performance in Financial Recommendations. AI & Society, 38(1), 89–102.

Singh, A., & Rao, V. (2023). Explainable AI for Banking Recommenders. AI Perspectives, 6(1), 9–28.

Gao, Y., & Wang, L. (2021). Big Data Analytics in Customer Personalization. Procedia Computer Science, 181, 1076–1085.

Ahmed, S. et al. (2020). Challenges and Solutions in AI-Driven Recommenders. Journal of Financial Transformation, 50, 34–45.

Johnson, K., & Mehta, S. (2020). Customer Profiling and Recommendation Engines. Decision Support Systems, 135, 113322.

Alzubi, J. A., & Nayyar, A. (2021). Smart Finance with AI Recommenders. AI Review, 54(3), 589–614.

URL: https://www.finextra.com/recommender-fintech (Visit Date: March 25, 2025)

URL: https://www.kaggle.com/datasets/bank-recommendation (Visit Date: March 25, 2025)

Li, P. et al. (2022). Banking Recommendations with XGBoost. Neurocomputing, 470, 196–205.

CHAPTER 8

AI BASED MOBILE ATTENDANCE SYSTEM WITH MULTI LAYERED SECURITY ARCHITECTURE

İLKER YILDIZ[1]

Introduction

Monitoring student attendance is not merely confined to recording physical presence in the classroom. Rather, it plays a critical role in tracking instructional accountability, maintaining academic discipline, and facilitating program evaluations based on learning outcomes. With the implementation of the Bologna Process, structural obligations such as institutional accountability, quality assurance, and transparency have compelled higher education institutions to monitor student attendance data both quantitatively and qualitatively in a regular and systematic manner.

However, many universities and secondary education institutions still rely on traditional procedures such as handwritten signature sheets, physical attendance forms, and manually updated logbooks supervised by instructors. These methods are subject to a

[1] Assistant Professor, Bolu Abant İzzet Baysal University, Bolu Technical Sciences Vocational School, Department of Computer Technologies, Orcid: 0000-0002-1575-2673, ilkeryildiz@ibu.edu.tr

range of systemic limitations, including the lack of identity verification, susceptibility to human error, operational inefficiency, and inadequate integration with digital infrastructures. Particularly in classrooms with high student density, such practices lead to significant time loss, compromise the validity and reliability of assessment protocols, expose critical vulnerabilities in data security, and enable ethically problematic behaviors such as proxy attendance.

Therefore, both the real-time verification of attendance and the systematic processing of the collected data necessitate a digital transformation that surpasses the limitations of traditional systems. Relying solely on physical presence or verbal declarations is no longer sufficient for accurate authentication in digital learning environments. Accordingly, the verification of student identity must be conducted through an integrated framework of multilayered digital security protocols, including biometric recognition, geographic location validation, liveness detection, device integrity analysis, and session monitoring. Within the context of institutional security architecture, such a comprehensive approach has become an indispensable requirement. The development of mobile, AI-driven, and multilayered authentication systems capable of addressing these needs in a secure and integrated manner has thus emerged as a strategic imperative in the landscape of contemporary educational technologies.

In this study, an artificial intelligence–based mobile attendance system equipped with multi-layered security mechanisms has been designed to ensure that student attendance in educational institutions is recorded in a more secure, efficient, rapid, and verifiable manner. The system comprises two separate applications: one for students and one for instructors. Developed using the Flutter framework and built on Google Firebase infrastructure, the proposed architecture integrates multiple layers of security within a single unified platform. These include QR code authentication, facial recognition via Google ML Kit, liveness detection, GPS-based location verification, repeat attempt control, time constraints, mock location detection, screenshot prevention, multi-factor user authentication via Firebase Authentication, and real-time data recording using the Firestore database. This architecture enables the identification of fraudulent users and ensures an end-to-end secure and auditable attendance process. The implementation of advanced security measures forms the foundation of the system's multi-dimensional verification model. While providing students with a personalized and secure authentication process, the system also equips instructors with a comprehensive absence reporting tool. Utilizing the Firebase infrastructure, all data are managed in the cloud environment, ensuring real-time synchronization, accessibility, and data integrity.

In the student application, the system permits each student to scan only their uniquely assigned QR code and prevents the reuse of the same code for subsequent attempts. Following this, the student is required to scan the dynamic QR code generated by the instructor specifically for the active course session. A facial image captured through the student's mobile device camera is then transmitted to the system's facial recognition module for identity verification. Subsequently, the student must pass a series of liveness detection tests, including eye blinking and head movement recognition. Upon successful biometric verification, the student's real-time geolocation data is transmitted to the system. The location validation algorithm verifies whether the student is within a 20-meter radius of the instructor's predefined coordinates. Students who complete all verification steps within the designated three-minute time frame are successfully marked as present in the attendance system. All processes are logged with a timestamp and user identifier via Firebase, ensuring traceability and auditability.

The instructor application provides functionalities such as course creation, student enrollment, listing enrolled students, initiating attendance sessions, and generating absenteeism reports. For each course session, the system dynamically generates a distinct QR code, which is transmitted to the student application in real time. Once the attendance session is initiated, the system automatically enforces a three-minute validity period, after which the session is terminated.

The reporting system analyzes student absenteeism on a weekly basis and presents the data through an intuitive, color-coded visual interface: green for 0 weeks of absence, yellow for 1–2 weeks, orange for 3–4 weeks, and red for 5 or more weeks. This interface offers instructors a clear and user-friendly representation of attendance trends and risks.

System Architecture and Technological Stack

The developed mobile attendance application is built upon a modular client–server software architecture. The system comprises two distinct client modules teacher and student applications both implemented using the Flutter framework. All data processing and identity verification operations are handled through the Google Firebase platform, which serves as the backend infrastructure. The application adopts a multi-layered security architecture that integrates QR code-based authentication, facial recognition, liveness detection, geolocation verification, mock location prevention, session time constraints, screen security, and systematic data logging. These mechanisms are orchestrated within a unified and coherent structure to ensure secure, reliable, and context-aware attendance verification.

Architectural Layers

The system is structured around five fundamental architectural layers, as described below:

1. **Presentation Layer**

This layer constitutes the user interface of the application, designed using Flutter's UI components. It encompasses page navigation, user notifications (SnackBars and Dialogs), adaptive design principles, and the implementation of Google's Material Design guidelines. Material Design serves as a primary reference framework, ensuring consistency and intuitiveness in terms of color schemes, typography, component arrangement, and user interactions (Google, 2023).Flutter's widget-based architecture enables a native and seamless user experience, while its adaptive layout capabilities facilitate responsive designs tailored to various device sizes and platform types (Wickramathilaka et al., 2025).

2. **Client-Side Logic Layer**

This layer is responsible for executing core client-side functionalities, including QR code generation and scanning, camera activation/deactivation control, facial detection, liveness analysis, and geolocation data processing.

Reactive data flow between UI components and underlying application logic is facilitated by the Riverpod state management framework, which ensures efficient and consistent state synchronization within the Flutter environment (Pinandito, 2023).

3. **Data and Identity Management Layer (Firebase Layer)**

This layer consists of cloud-based services that manage the core data security infrastructure and user authentication processes of the mobile attendance system. Key functionalities such as managing student and teacher identities, session-based data operations, and secure storage of media content are executed within this architectural layer.

Firebase Authentication is employed to verify the identities of users logging into the application. The authentication process supports both password-based and token-based mechanisms, enabling secure management of session states (Firebase, 2024a).

Cloud Firestore serves as a scalable, low-latency NoSQL database in which all structural data such as students, courses, sessions, and attendance records are hierarchically organized. Firestore ensures real-time synchronization of data changes initiated from the client side, thereby maintaining data integrity and concurrency (Firebase, 2024b).

Firebase Storage provides cloud-based storage specifically optimized for securely maintaining media files, such as biometric images. The system supports high-performance transfer of large files while restricting access to authorized devices only (Firebase, 2024c).

Firebase Security Rules enforce granular access controls to the database and file system based on user identity and application context. These rules constitute a critical security layer, effectively mitigating client-side tampering and unauthorized data access (Firebase, 2024d).

Firebase App Check validates whether incoming requests originate from a verified and unmodified instance of the client application. This mechanism defends against spoofed apps, tampered clients, and unauthorized access attempts. Security-critical operations—such as identity verification and data retrieval—are not executed unless App Check validation is passed (Firebase, 2024e).

4. **Security Layer**

This layer encompasses security mechanisms designed to protect application operations on the client side without disrupting the user experience. It ensures the protection of on-screen content against unauthorized access, enforces session time constraints, and implements fine-grained access control over data resources.

5. **Monitoring & Logging Layer**

This layer is designed to continuously monitor and log the system's operational history, user behaviors, and security-related events in real time. It evaluates critical occurrences such as on-time attendance submission, geolocation accuracy, multi-device login attempts, and the temporal validity of scanned QR codes.

Using Cloud Firestore Streams**,** student sessions are tracked instantaneously. Each transaction is recorded in the database along with its associated timestamp, user identifier, and session parameters. This infrastructure enables algorithmic absenteeism analysis within the teacher application. The resulting data is presented in the form of graphical reports and visual dashboards to facilitate actionable insights (Google, 2024f).

Security Layers: Multi-Factor Authentication Approach

1. QR Code Verification and Replay Prevention

Advancements in image processing algorithms and the enhanced optical resolution of mobile device cameras have enabled QR codes to be scanned with high accuracy and sub-second latency from any 360° orientation. This capability significantly enhances both user experience and system efficiency in mobile applications requiring fast and contactless identity verification (Liew et al., 2021).

In the proposed system, QR codes are structured to serve two distinct verification phases: student identity authentication and session-based attendance verification**.**

QR Code Generation in the Teacher Application

Student Specific QR Codes: Unique QR codes are generated for each student during the initial registration phase for a given course and stored under the respective lesson node in the Firestore database. These codes are dynamically created and contain encoded identity data such as the student's full name and student number, as well as system-level identifiers including `studentId`, `lessonId`, and `sessionId`. These parameters enable precise, time-bound verification, ensuring each QR code is valid only for the designated student and session. To prevent replay attacks, the system tracks QR usage locally via `SharedPreferences`.

Session-Specific QR Code: When initiating an attendance session, the teacher dynamically generates a QR code through the system interface. This code embeds session-level metadata including `sessionId`, `lessonId`, `timestamp`, and a predefined validity window (3 minutes). This ephemeral code functions as a one-time, time-sensitive authentication token, adding an additional temporal security layer to each session.

Student Application – QR Code Scanning and Verification Process

In the student application, the user initiates the identity verification process by scanning their system-specific QR code. Subsequently, the user scans the session-specific QR code generated by the instructor at that moment.

Following this process, the application performs several checks on Firebase Firestore: it verifies the existence and validity of the session associated with the `sessionId`, confirms whether the student is enrolled in the corresponding course using the `studentId`, and checks whether the same QR code has already been scanned.

If the verification is successful, the student proceeds to the next security layer of the attendance application. Otherwise, the system automatically rejects repeated or invalid attempts. This structure ensures that both identity-based and time-based security validations are enforced simultaneously through a multi-layered authentication approach.

2. Face Recognition

Face recognition systems generally operate in three sequential stages: face detection, feature extraction, and comparison. In modern applications, these processes are typically implemented using convolutional neural networks (CNNs) based on deep learning architectures. Mobile-compatible solutions such as Google ML Kit offer pre-trained models that efficiently perform these three stages in an optimized manner (Google ML Kit, 2025).

Using Google ML Kit, the student's real-time face image is compared against the corresponding reference image stored in the database. Face feature vectors are extracted, and cosine similarity is computed between them (matching threshold ≥ 0.6).

3. Liveness Detection

Liveness detection, a critical layer of security in mobile face verification systems, aims not only to detect the presence of a face but also to determine whether the detected face represents a live, real-time, and interactive biometric source. In this study, Google ML Kit's face detection API is utilized to analyze user blink patterns, head orientation (Euler angles), and dynamic changes in facial landmarks (such as eyes, mouth, and nose).

During the liveness detection process, the incoming video stream from the user is analyzed frame by frame. For each frame, the *eye open probability* is computed to determine whether the eyes are open or closed. Simultaneously, Euler angles (X, Y, Z axes) are tracked to assess head movements. Facial landmarks such as the positions of the eyes, mouth, and nose are monitored over time. These features are evaluated collectively to infer the liveliness of the detected face (Breuel, 2021).

This method goes beyond static face detection by analyzing real-time biometric interaction, thereby enabling robust user verification. Consequently, the system becomes more resilient to face-based spoofing attacks.

In this implementation, both blinking and head motion are monitored. Blink detection is performed using the Eye Aspect Ratio (EAR) method.

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|} \tag{1}$$

Here, p_1 and p_4 represent the outer corner points along the horizontal axis of the eye. p_2, p_3, p_5 $andp_6$ correspond to the upper and lower landmark points along the vertical axis of the eye. The notation $\|.\|$denotes the Euclidean distance between two points (Dewi et al., 2022).If the EAR (Eye Aspect Ratio) value is high, the eye is considered open. If the EAR value drops below a certain threshold, the eye is considered closed.

Head pose estimation (yaw/pitch > ±15°) is analyzed for head movement.

$$s.\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K.[R|t]..\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2}$$

Here, (u, v) represent 2D landmark points on the image plane (e.g., eye corners, nose tip, mouth).

(X, Y, Z) denote the corresponding 3D model coordinates (i.e., real-world reference points on the face).

K is the camera intrinsic parameter matrix.

R is the rotation matrix (Euler angles: yaw, pitch, roll).

t is the translation vector, representing the position of the head relative to the camera.

s is the scale factor.

This formula is used to compute the head's orientation in three-dimensional space (Zeng et al., 2023). The resulting Euler angles are interpreted as follows:

Yaw (Z-axis) → Left–right rotation

Pitch (X-axis) → Up–down tilting

Roll (Y-axis) → Side tilting (head turning)

4. Location Verification (GPS)

The location verification module, which constitutes an additional security layer in the mobile attendance system, is designed to ensure that students can participate in attendance only within predefined physical boundaries. Prior to initiating the attendance process, students are required to share their device's GPS data with the system. This location data is compared with the instructor's predefined location for the corresponding class session, and the distance between the student and teacher locations is calculated using the Haversine algorithm in eq 3 (Ikasari & Widiatuti, 2021). If the calculated distance exceeds 20 meters, the verification is considered invalid and the attendance process is terminated.

This structure not only guarantees physical presence but also serves as a multi-layered defense mechanism against deceptive system manipulations such as mock location usage. Accordingly, the system actively monitors two technical control points on Android-based devices:

isFromMockProvider: Detects whether the location information is derived from a mock provider.

Settings.Secure.ALLOW_MOCK_LOCATION: Checks whether the mock location feature is enabled on the device.

$$a = sin^2\left(\frac{\Delta_\emptyset}{2}\right) + \cos(\emptyset_1)\,.\,cos(\emptyset_2).\,sin^2\left(\frac{\Delta_\lambda}{2}\right) \qquad (3)$$

$$c = 2.\operatorname{arctan} 2\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R.c$$

Here, $\emptyset_1$ *and* $\emptyset_2$ represent the latitudes of the teacher and the student, respectively (in radians).

$\Delta_\emptyset$ is the difference in latitude (in radians).

Δ_λ is the difference in longitude (in radians).

R denotes the Earth's radius (approximately 6,371 km).

d is the distance between the two points (in km or m).

5. Mock Location Detection

The use of mock locations can be detected in the Flutter environment through libraries such as `detect_fake_location` and `safe_device`. On Android, methods like `position.isMocked` are utilized, while on iOS, detection typically involves jailbreak-related checks (Fazzini et al., 2021; Zeexan, 2025).

6. Time Limit and Timer Control

Once attendance is initiated, all related processes must be completed within 180 seconds. This constraint ensures session continuity and increases resistance to fraudulent activities. A timer generated via Cloud Functions automatically terminates the attendance session upon timeout.

7. Screenshot and Screen Recording Prevention

In mobile attendance applications, securing digital content requires specific measures against external access threats that may compromise system integrity.

Android: The application screen is flagged using Android's `FLAG_SECURE` attribute. This flag prevents screenshots and screen recordings, and also blocks content from being displayed via remote display protocols (Android Developers, n.d.; Google Play Console, n.d.).

iOS: The screen capture status is monitored using the `UIScreen.isCaptured` API. When screen recording is detected, the application can trigger content protection strategies such as obfuscation or blackout modes (Apple Developer, 2017; Medium, 2024).

8. Device Identity Binding

In the developed system, beyond verifying the user's identity, the device used by the student is also recognized and authenticated by the system. This ensures that only authorized devices are permitted to interact with the application, effectively restricting access from multiple devices at the technical level.

On the Android platform, a unique identifier specific to each device ANDROID_ID is retrieved via `Settings.Secure.ANDROID_ID`. This identifier is considered a long-lived and reliable reference, remaining constant unless the device is factory reset (Google Developers, 2024). Upon the first launch of the application, this device ID is captured and stored in Firebase, linked to the user's identity information. If a student attempts to log in from a different device using the same account, the system compares the new device's `ANDROID_ID` with the previously registered value. If the match fails, the system notifies the user or restricts access accordingly.

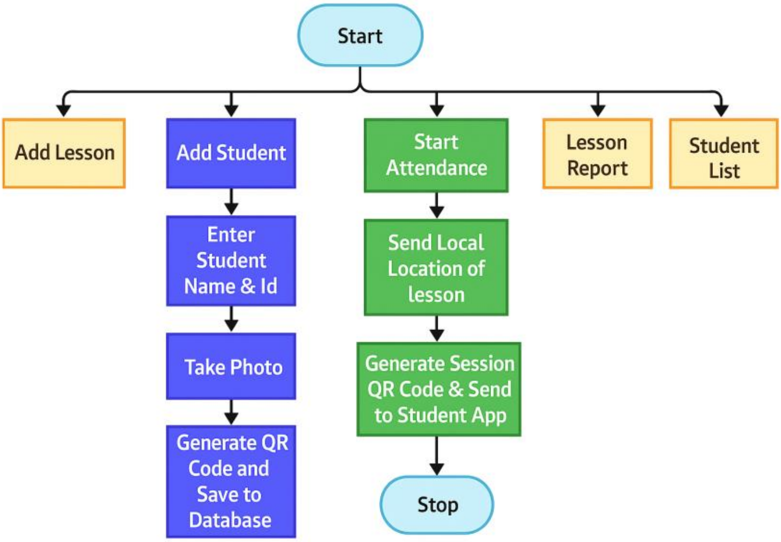
On iOS, this verification process is implemented using Apple's Keychain service (Apple Developer, 2023).

In Flutter applications, this functionality is integrated through the `flutter_secure_storage` package, enabling secure device identification and verification across both platforms.

System Architecture

The flowcharts below illustrate the fundamental process steps and data flows of the teacher and student applications in a logical sequence. Figure 1 presents the process flow of the teacher application, while Figure 2 depicts the corresponding flow for the student application.

Figure 1. General process flow diagram of the teacher application.

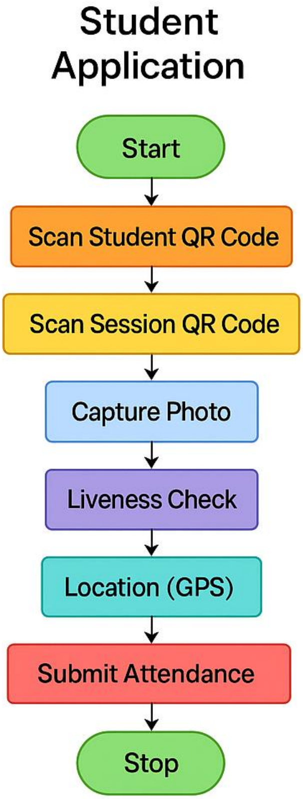

Start
Add Lesson
Add Student
Start Attendance
Lesson Report
Student List
Enter Student Name & Id
Send Local Location of lesson
Take Photo
Generate Session QR Code & Send to Student App
Generate QR Code and Save to Database
Stop

In the teacher application, the instructor is able to perform operations such as adding lessons, registering students, initiating attendance sessions, retrieving attendance reports, and viewing student lists for each course. The process begins with **Add Lesson**, followed by the entry of student information (e.g., name and surname) and real-time photo capture.

For each student, a unique QR code is generated using this data and stored in the database.

When the attendance session is initiated (Start Attendance), the teacher's device transmits its current geolocation to the system. A dynamic, session-specific QR code is then generated and distributed to the student application. Additionally, the teacher can access Lesson Reports and Student Lists**.**

Figure 2. General process flow diagram of the student application.



The student application begins with a two-step QR code scanning process: first, the student scans their personalized QR code, followed by the session-specific QR code generated for the current class. In the next step, the student captures a real-time photo, which is processed by the system to be compared against the reference image previously stored by the teacher.

Subsequently, liveness detection is performed by analyzing facial dynamics such as blinking and head movements to verify that the face is genuine and belongs to a live individual. During the location verification phase, the student's physical location is retrieved and compared with the predefined teacher location for the session.

Once all verification steps are successfully completed, the student submits their attendance. This submission is recorded along with a timestamp and device metadata. If an attempt is made to access the system from a device other than the one used during the initial identity verification, the system denies access. Each user's device identifier (device ID) is stored in the database during the first session, and all subsequent sessions require this identifier to be validated, allowing access only from authorized devices.

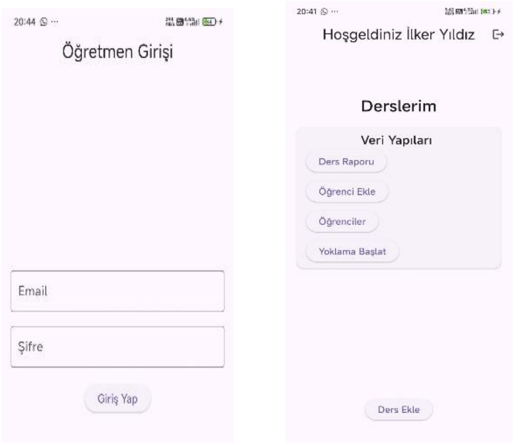
Designed Mobile Attendance Applications

The application architecture is built on modular and reusable components, taking into account both user experience (UX) and security requirements. Screenshots and descriptions of the designed applications used in this study are provided below.

Teacher Application

The teacher application provides a management interface through which instructors can perform tasks such as adding courses, registering students, viewing student lists for each course, initiating attendance sessions, and analyzing absenteeism data. The main screen and post-login interface of the teacher application are shown in Figure 3.

Figure 3. Main screen of the teacher application.

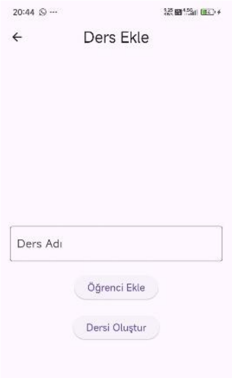


The core modules of the application are described below.

User Authentication and Role-Based Access Control: Teacher identity verification is performed using Firebase Authentication, with role management implemented via customClaims. Each teacher is granted access exclusively to the lessons and student data they have created.

Lesson Addition: During the lesson addition process, the lesson name is collected and stored in a lessons collection within Firestore. Each lesson document contains a reference to the teacher's UID. A screenshot of the lesson addition process is shown below.

Figure 4. Screenshot of the lesson addition screen in the teacher application.



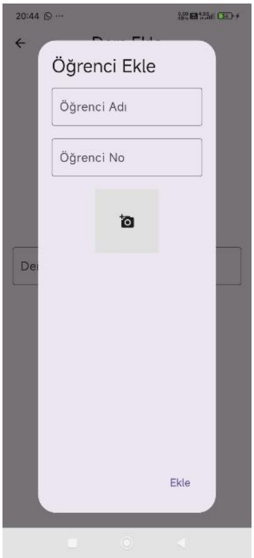
Student Registration, Photo Management, and QR Code Generation (Student): A user record is created in the system by capturing the student's name, ID number, and facial photograph. After obtaining the student's facial image via the device camera, the Flutter application converts it into a byte array and stores it in Firebase Storage. Storing the data in byte format optimizes both data size and access speed. Furthermore, this approach is advantageous for reprocessing the image, as it can be converted into different resolutions and formats as needed.

Firebase Storage leverages Google Cloud infrastructure to securely store media files, with access control enforced via Firebase Security Rules. This ensures that only authorized users can access the image data.

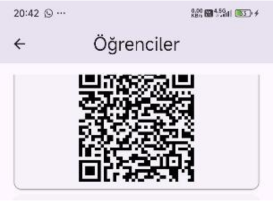
Simultaneously, the visual data is processed by Google ML Kit to extract facial embeddings, which serve as references for subsequent face recognition and verification tasks. Thus, the system achieves a secure and highly accurate student identity verification process.

Additionally, a unique QR code is generated for the registered student and saved in the database. A screenshot of the student registration screen is provided in Figure 5.

Figure 5. Screenshot of the student registration screen in the teacher application.

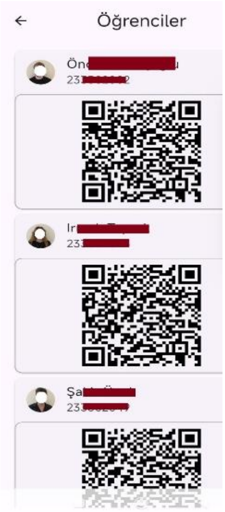


After entering the student information, the personalized QR code generated for the student is displayed in the screenshot shown in Figure 6.



Students: The names, surnames, student numbers, photographs, and personalized QR codes of students registered for the course can be viewed here. The corresponding screen is shown in Figure 7.

Figure 7. List of students enrolled in the course.



Öğrenciler

Ön[redacted]
23[redacted]

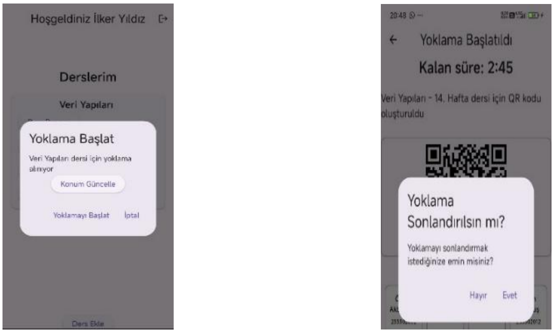
Ir[redacted]
23[redacted]

Şa[redacted]
23[redacted]

Attendance Session QR Code Generation (Participation Session): During the weekly attendance session for a course, the teacher transmits their current physical location via GPS. Based on this location, a dynamic QR code valid only for that specific course session is generated and stored in the attendance_sessions collection with a validity period of 3 minutes.

A screenshot of the attendance initiation process is shown in Figure 8. Additionally, the teacher has the ability to terminate the attendance session early if necessary.

Figure 8. Attendance session initiation screen for the course.



Hoşgeldiniz İlker Yıldız

Derslerim

Veri Yapıları

Yoklama Başlat

Veri Yapıları dersi için yoklama alınıyor

Konum Güncelle

Yoklamayı Başlat İptal

Ders Ekle

20:48

Yoklama Başlatıldı

Kalan süre: 2:45

Veri Yapıları - 14. Hafta dersi için QR kodu oluşturuldu

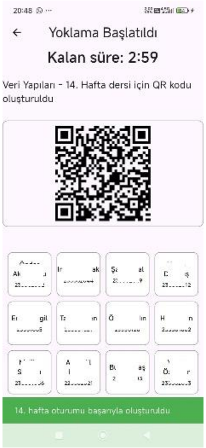
Yoklama Sonlandırılsın mı?

Yoklamayı sonlandırmak istediğinize emin misiniz?

Hayır Evet

When the Start Attendance button is pressed, the session-specific QR code is generated and the attendance process begins. The session QR code is shown in Figure 9.

Figure 9. QR code for the course session.



Yoklama Başlatıldı

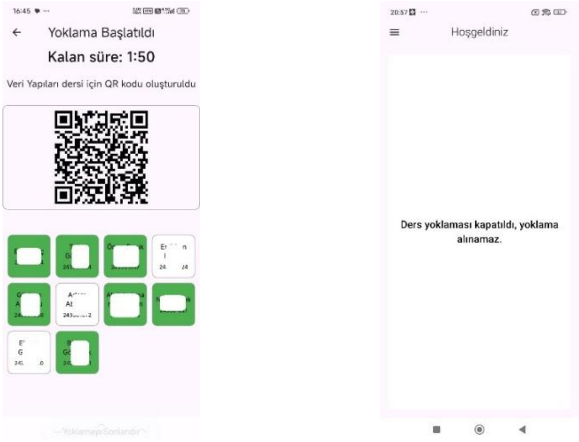
Kalan süre: 2:59

Veri Yapıları - 14. Hafta dersi için QR kodu oluşturuldu

14. hafta oturumu başarıyla oluşturuldu

Attendance Duration Tracking: A timer automatically initiated by Cloud Functions invalidates the QR code and closes the attendance session after a 3-minute period. Students who have completed attendance are displayed in green on the screen. Attendance process screenshots are shown in Figure 10.

Figure 10. Attendance process interface.



Yoklama Başlatıldı

Kalan süre: 1:50

Veri Yapıları dersi için QR kodu oluşturuldu

Yoklamayı Sonlandır

Hoşgeldiniz

Ders yoklaması kapatıldı, yoklama alınamaz.

Absenteeism Reports and Visualization: Attendance data retrieved from Firebase is analyzed to calculate absentee counts per student, with status indicated graphically using a color-coded scheme (green–yellow–orange–red). Weekly attendance results are reported on a weekly basis. Students who do not attend a session are marked in red as "Absent," while those who attend are marked in green as "Present." Additionally, each session displays the total number of students, along with counts of those present and absent.

Categorical Classification of Absenteeism Status

Student performance is evaluated not only based on attendance but also considering attendance consistency over time. Each course encompasses a 14-week instructional period according to the institution's curriculum. Absenteeism status is calculated across these 14 sessions. The categorical classification is presented in Table 1.

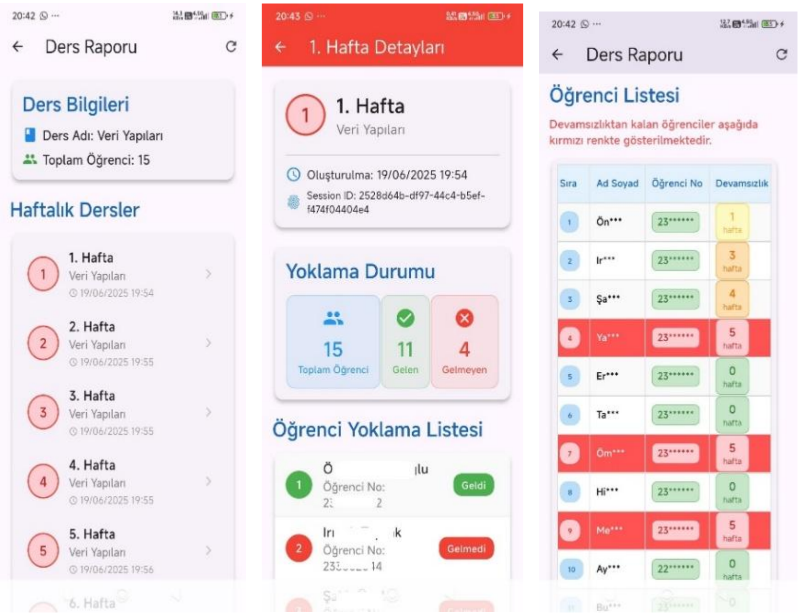
Table 1. Categorical Classification of Absenteeism Status

Absenteeism Week (d)	Category	Color	Academic Description
$d = 0$	Excellent	Green	Full attendance recorded.
$1 \leq d \leq 2$	Good	Yellow	Low-level absenteeism recorded
$3 \leq d \leq 4$	At-Risk	Orange	Intervention required.
$d \geq 5$	Critical	Red	Failed due to excessive absenteeism.

Color categories are presented alongside student information in the teacher application's reporting module to facilitate intuitive decision-making.

Since a university course typically spans 14 weeks, a total of 14 sessions are created per course. Upon completion of the 14 weeks, students who fail due to absenteeism are listed in red. The reporting interface is shown in Figure 11.

Figure 11. Reporting interface screenshots.



Ders Raporu

Ders Bilgileri

Ders Adı: Veri Yapıları

Toplam Öğrenci: 15

Haftalık Dersler

1. Hafta
Veri Yapıları
19/06/2025 19:54
2. Hafta
Veri Yapıları
19/06/2025 19:55
3. Hafta
Veri Yapıları
19/06/2025 19:55
4. Hafta
Veri Yapıları
19/06/2025 19:55
5. Hafta
Veri Yapıları
19/06/2025 19:56
6. Hafta

1. Hafta Detayları

1 1. Hafta
Veri Yapıları

Oluşturulma: 19/06/2025 19:54

Session ID: 2528d64b-df97-44c4-b5ef-f474f04404e4

Yoklama Durumu

15 Toplam Öğrenci | 11 Gelen | 4 Gelmeyen

Öğrenci Yoklama Listesi

1 Ö ılu
Öğrenci No: 2: 2 — Geldi

2 Irı k
Öğrenci No: 23 14 — Gelmedi

Şa

Ders Raporu

Öğrenci Listesi

Devamsızlıktan kalan öğrenciler aşağıda kırmızı renkte gösterilmektedir.

Sıra	Ad Soyad	Öğrenci No	Devamsızlık
1	Ön***	23******	1 hafta
2	Ir***	23******	3 hafta
3	Şa***	23******	4 hafta
4	Ya***	23******	5 hafta
5	Er***	23******	0 hafta
6	Ta***	23******	0 hafta
7	Öm***	23******	5 hafta
8	Hi***	23******	0 hafta
9	Me***	23******	5 hafta
10	Ay***	22******	0 hafta
11	Bu***	23******	0

Student Application

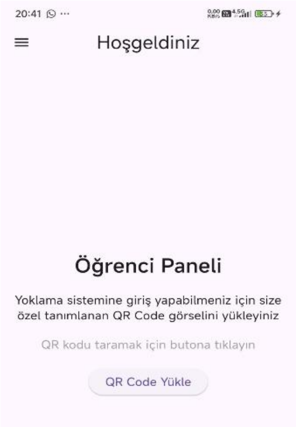
The student application offers a highly secure multi-layered authentication process to ensure student participation in assigned courses. The process is structured as follows:

Student Identity Verification and QR Code Scanning: Upon launching the application, the student can scan their unique QR code generated by the teacher specifically for their account. The system verifies whether the QR code has been previously scanned. If the QR code was already used, access to the system is denied. If the QR code is valid, the student is welcomed with a “Welcome” message along with the date and time of access.

User identity is authenticated using QR code data via Firebase Authentication. For each session, a JWT (JSON Web Token) is generated and utilized as a temporary and secure identity token for all client–server communications.

A screenshot of the student panel is provided in Figure 12.

Figure 12. Student panel.



20:41

Hoşgeldiniz

Öğrenci Paneli

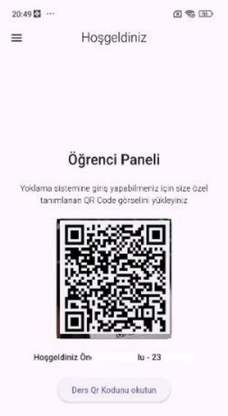
Yoklama sistemine giriş yapabilmeniz için size özel tanımlanan QR Code görselini yükleyiniz

QR kodu taramak için butona tıklayın

QR Code Yükle

Course Session Attendance QR Code Scanning: The student scans the QR code corresponding to the session currently initiated by the teacher. This QR code is valid only within its designated time window. If the QR code has been used previously, the system denies attendance access. For each session, a JWT (JSON Web Token) is generated and utilized as a temporary and secure identity token for all client–server transactions. The course session attendance panel is shown in Figure 13.

Figure 13. Course session attendance panel.



20:49

Hoşgeldiniz

Öğrenci Paneli

Yoklama sistemine giriş yapabilmeniz için size özel tanımlanan QR Code görselini yükleyiniz

Hoşgeldiniz Ön[redacted]lu - 23[redacted]

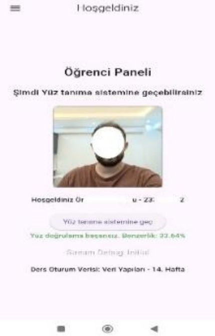
Ders Qr Kodunu okutun

Face Verification via Camera: The application requests camera access to capture a live image of the student. Face detection is performed using the Google ML Kit Face Detection API. A 128-dimensional facial embedding extracted from the student's previously recorded image is compared with the embedding derived from the live image using the cosine similarity metric. If the similarity score exceeds 0.60, identity verification is considered successful. Image preprocessing techniques are applied during this process to enhance recognition accuracy.

Face Recognition Verification Failure: The developed system only permits progression to the next stage if real-time visual verification is associated with the correct student identity. If an image not belonging to the student (e.g., a screenshot of another person, a passport photo, or another device's photo) is presented to the system camera, the face recognition algorithm invalidates the identity verification due to low similarity scores and halts the operation flow.

Figure 14 depicts a scenario where a user has logged in via QR code, but subsequently an image of a different individual is presented to the system. The face recognition module provided by Google ML Kit detected a similarity score below 60%, marking the verification step as "failed."

Figure 14. Face recognition.



Öğrenci Paneli

Şimdi Yüz tanıma sistemine geçebilirsiniz

Hosgeldiniz Ör u - 23 2

Yuz doğrulama başarısız. Benzerlik: 33.64%

Sistem Debug: Initial

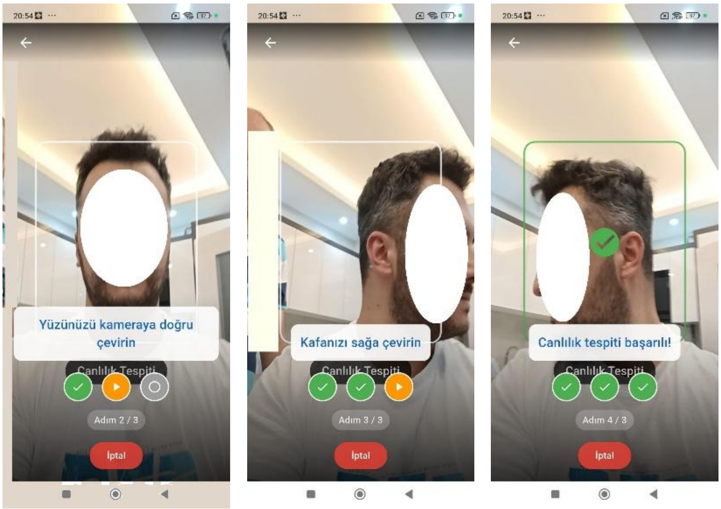
Ders Oturum Verisi: Veri Yapıları - 14. Hafta

Liveness Detection: To prevent spoofing attempts using static images or videos during face recognition, a real-time liveness detection module has been integrated. The user is required to perform blink actions, which are analyzed using the Eye Aspect Ratio (EAR) metric. Subsequently, the student is asked to turn their head to the right and left. Head movement (yaw/pitch) is detected through pose estimation. If the angular change between consecutive frames exceeds ±15°, liveness is confirmed. This mechanism prevents operations using fake faces or screenshots. The application mandates that these tests be completed within 10 seconds.

Designed to enhance the biometric authentication security of the attendance system, the liveness detection module requests the user to perform specific dynamic facial actions in real time. These actions sequentially include blinking and head rotations (turning right or left). The system continuously monitors these movements visually and via facial landmarks identified by the Google ML Kit Face Detection API. Each action is logged synchronously, and upon reaching predefined threshold values, liveness verification is deemed "successful."

Liveness detection constitutes a security layer specifically aimed at preventing the use of static images (such as screenshots or printed photos) within the system. The system's responsiveness exclusively to dynamic, voluntary facial movements substantially mitigates spoofing attempts. In this respect, the liveness detection algorithm ensures the mobile attendance system's robustness against spoofing attacks and significantly enhances the security of the identity verification process.

Figure 15. Stages of liveness detection.



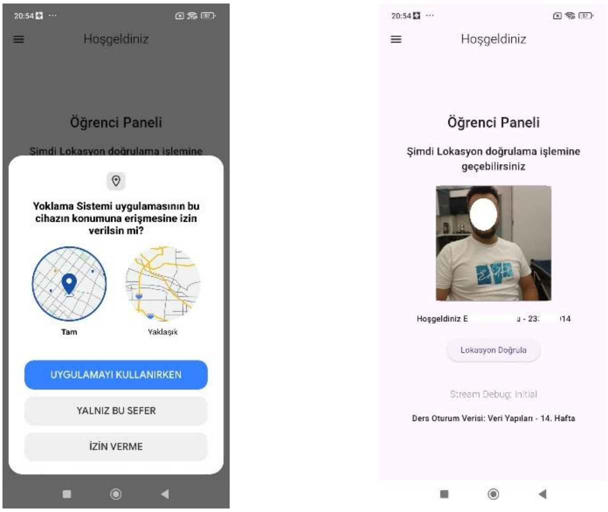
Yüzünüzü kameraya doğru çevirin

Kafanızı sağa çevirin

Canlılık tespiti başarılı!

Location Verification and Mock Location Detection: Screenshots of the location verification process in the student application are presented in Figure 16.

Figure 16. Location verification.



Öğrenci Paneli

Şimdi Lokasyon doğrulama işlemine

Yoklama Sistemi uygulamasının bu cihazın konumuna erişmesine izin verilsin mi?

Şimdi Lokasyon doğrulama işlemine geçebilirsiniz

Hoşgeldiniz E ı - 23: 14

Stream Debug: Initial

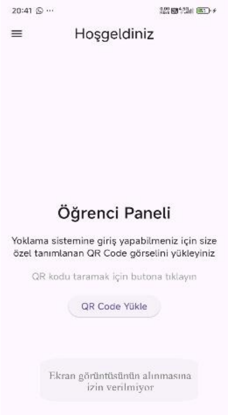
Ders Oturum Verisi: Veri Yapıları - 14. Hafta

Time Limit and Session Duration Control: The QR code generated by the teacher for the session is valid for only 3 minutes.

Requests exceeding this time limit are rejected through timestamp verification. The attendance session is closed using a timer created by Cloud Functions.

Screen Security and Prevention of Unauthorized Visual Access: When a user attempts to capture a screenshot while the application is running, the system immediately detects this action and notifies the user with a visual warning message. This intervention not only prevents the capture but also serves as a deterrent by communicating the system's security policies to the user. Similarly, content duplication is blocked in cases where screen recording is initiated or third-party monitoring applications run in the background. Figure 17 presents a user interface example demonstrating the security layer's prevention of screenshot capture.

Figure 17. Screenshot prevention interface.



Hoşgeldiniz

Öğrenci Paneli

Yoklama sistemine giriş yapabilmeniz için size özel tanımlanan QR Code görselini yükleyiniz

QR kodu taramak için butona tıklayın

Ekran görüntüsünün alınmasına izin verilmiyor

Completion of Attendance and Data Recording: The developed mobile attendance application concludes with an automated recording step that registers users who successfully complete the multi-layered security verification process. This verification sequence is initiated after the successful completion of consecutive security protocols including QR code scanning, face recognition, liveness detection, location verification, and replay prevention. These layers collectively aim to authenticate the user’s identity and physical presence synchronously through multi-factor verification.

The system prohibits progression to attendance recording unless all verification conditions are met, thereby technically eliminating risks of client-side fraud and incomplete data submission.

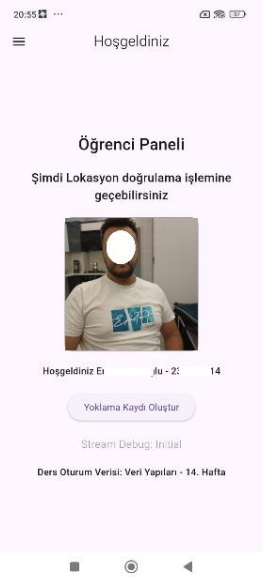
Once the student successfully passes all security stages, a “Submit Attendance” button becomes active in the application interface. Upon pressing this button, the client initiates data transfer to the cloud database configured on Firebase Firestore using a CreateOrUpdate methodology. During this process, the system instantaneously records student-specific unique identifiers (UID), dynamic session IDs, attendance timestamps (in UTC format), geolocation data (latitude, longitude, accuracy), and device session information (device ID, operating system, application version).

This information is transmitted to both Firestore and, simultaneously, to Firebase Realtime Database or Cloud Logging services, ensuring full traceability of the attendance transaction. Thus, not only student participation but also session metadata are digitally recorded.

Furthermore, the data is logged into system audit trails for retrospective inspection and is displayed in real-time reports on the teacher application panel. This approach facilitates the attendance process to be monitored, queried, and audited transparently and historically.

The user interface design and operational step corresponding to this final stage of the application are presented in Figure 18. The screenshot reflects the state when the student gains access to the “Submit Attendance” button and visually represents the system’s final verification step based on user interaction.

Figure 18. Attendance submission screen activated upon successful completion of all verification layers.



Öğrenci Paneli

Şimdi Lokasyon doğrulama işlemine geçebilirsiniz

Hoşgeldiniz E[redacted]lu - 2[redacted]14

Stream Debug: Initial

Ders Oturum Verisi: Veri Yapıları - 14. Hafta

Resistance to Attack Scenarios

Mobile attendance systems are vulnerable to various security threats, including identity fraud, replay attacks involving QR codes, spoofed location reports, and screen recording. Therefore, the developed system incorporates multi-layered security measures designed to counteract potential attack scenarios. Table 2 systematically presents possible attack scenarios alongside the corresponding defense mechanisms implemented to mitigate these threats.

Table 2. Implemented Security Countermeasures for Potential Attack Vectors

Attacks	Implemented Countermeasures
Participation using QR code screenshot	QR code reuse prevention, user-to-code binding and validation
Face recognition spoofing via static image	Liveness detection using blink and head movement recognition
Location spoofing via fake GPS	Mock location detection combined with 20-meter proximity validation
Screen recording or screenshot capture	Android FLAG_SECURE
Participation outside allowed time window	3-minute countdown timer and server-side timestamp validation
Unregistered device access	Access restricted to pre-authorized device only

Usability and System Evaluation

The system design is based on the usability principles outlined in ISO 9241-210. Although formal user testing has not yet been conducted in field environments, system behavior has been assessed through simulated analyses.

Simulated Test Results

The student application, developed on the Flutter framework, is structured to guide the attendance process step-by-step. The sequence of operations includes: scanning the student-specific QR code, identifying the course session code, performing face recognition (using Google ML Kit Face Detection API), executing liveness detection (blink and head movement), verifying location, and recording data to Firebase. All steps were measured using Firebase Performance Monitoring, and the resulting latency data were analyzed.

Simulated tests indicate that the system completes all operations in an average of 34 seconds (34,000 ms). This duration corresponds to approximately 18.9% of the defined 180-second operational limit, demonstrating that the application performs efficiently in terms of timing. Average durations for each operational step are presented in Table 3 below:

Table 3.. Process Steps and Average Latency Durations

Process Step	Average Duration(ms)	Execution Status
QR Code Scanning	2600	Successful
Face Recognition (Google ML Kit API)	8300	Successful
Liveness Detection (Blinking & Head Movement)	14200	Successful
Firebase Database Write Operation	3750	Successful
Firestore Report Query	2300	Successful

Latency analysis revealed that the processing time for liveness detection and face recognition modules is higher compared to other steps. The next phase of the system development involves conducting field-based usability testing with different user groups (teachers and students) across various device types and network environments.

Findings from real-world testing will provide concrete evidence of the system's usability level and will support sustainable improvement processes based on user feedback. In its current state, the system's operational steps, security procedures, and interface designs functionally operate and demonstrate a high potential for usability.

Conclusion and Recommendations

This study designed and implemented a mobile attendance system featuring a multi-layered security architecture supported by artificial intelligence, aiming to securely, rapidly, and verifiably record student attendance. Developed with Flutter and utilizing the Google Firebase infrastructure, the system consists of two separate mobile applications: one for teachers and one for students.

The proposed architecture integrates multiple security layers within a cloud-based framework, including QR code verification, face recognition (Google ML Kit), liveness detection (blink and head movement), GPS-based location verification, device control (device-specific operation permissions), mock location detection, replay prevention, session time limitations, screenshot prevention, multi-factor user authentication, and real-time data logging via Firebase Firestore.

Tests demonstrated that the entire operational process completes within an average duration of 3 minutes, with each transaction logged by Firebase alongside timestamps and user identities. The teacher application enhances user experience through functions such as session-specific QR code generation, student identification, and absenteeism reporting.

In conclusion, this work presents a modular, scalable, and highly secure attendance system that unifies biometric, geographic, and temporal verification methods within a singular framework. This architecture offers a valuable reference for researchers and institutions aiming to develop AI-supported secure mobile applications.

While the current study provides a comprehensive foundation for designing mobile-based secure attendance systems, future work plans include enhancing face recognition algorithms, implementing blockchain-based attendance logging, privacy-preserving biometric verification, and advanced anomaly detection methods.

Acknowledgements : The author extends their appreciation to Computer Programmer Ö. Kaan Sağlam for his valuable support in the programming phase of the Flutter-based mobile application implemented in this study.

References

Apple Developer. (2023). Keychain Services. Apple Developer Documentation. https://developer.apple.com/documentation/security/keychain_services

Apple Developer. (2017). Responding to screen capture in iOS 11 (Technical Q&A QA1970). Retrieved April 2025, from https://developer.apple.com/library/archive/qa/qa1970/_index.html developer.apple.com+1developer.apple.com+1

Android Developers. (2024). FLAG_SECURE | Android Developers documentation. Retrieved from https://developer.android.com/reference/android/view/WindowManager.LayoutParams#FLAG_SECURE

Andoid Developers. (n.d.). Secure sensitive activities using FLAG_SECURE. Retrieved April 2025, from https://developer.android.com/security/fraud-prevention/activities support.google.com+3developer.android.com+3learn.microsoft.com+3developer.apple.compub.dev+1stackoverflow.com+1

Android Developers. (2024). FLAG_SECURE | Protecting content. Retrieved from https://developer.android.com/reference/android/view/WindowManager.LayoutParams#FLAG_SECURE

Dewi, C., Chen, R. C., Jiang, X., & Yu, H. (2022). Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. PeerJ Computer Science, 8, e943.

Fazzini, M., Gorla, A., & Orso, A. (2020, December). A framework for automated test mocking of mobile apps. In Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (pp. 1204-1208).

Firebase. (2024a). Firebase Authentication documentation. Retrieved from https://firebase.google.com/docs/auth

Firebase. (2024b). Cloud Firestore documentation. Retrieved from https://firebase.google.com/docs/firestore

Firebase. (2024c). Cloud Storage documentation. Retrieved from https://firebase.google.com/docs/storage

Firebase. (2024d). Security Rules documentation. Retrieved from https://firebase.google.com/docs/rules

Firebase. (2024e). App Check documentation. Retrieved from https://firebase.google.com/docs/app-check

Google. (2024f). FlutterFire documentation – Cloud Firestore streams. Retrieved from https://firebase.flutter.dev/docs/firestore/usage/

Google. (2024g). Cloud Functions for Firebase. Retrieved from https://firebase.google.com/docs/functions

Google Play Console. (n.d). Protect your app and fight abuse with security flags FLAG_SECURE. Retrieved April 2025, from https://support.google.com/googleplay/android-developer/answer/14638385?hl=en support.google.com

Google Developers. (2024). Settings.Secure.ANDROID_ID. Android Developers. https://developer.android.com/reference/android/provider/Settings.Secure#ANDROID_ID

Google. (2025). ML Kit: Machine learning for mobile developers. https://developers.google.com/ml-kit

Google. (2023). Material Design guidelines. Material.io. Retrieved from https://m3.material.io

Ikasari, D., & Widiatuti, R. A. (2021). Implementation of Haversine Formula to Determine the Shortest Path Using Web Based Application for a Case Study of High School Zoning in Depok. American Journal of Software Engineering and Applications, 10(2), 19-31.

Jatnika, A. A. D., Akbar, M. A., & Pinandito, A. (2023). Comparative Analysis of the Use of State Management in E-commerce Marketplace Applications Using the Flutter Framework. Journal of Information Technology and Computer Science, 8(2), 111-124.

Liew, K. J., & Tan, T. H. (2021). QR code-based student attendance system. In 2021 2nd Asia Conference on Computers and Communications (ACCC) (pp. 10-14). IEEE.

Medium. (2024). Preventing app from screen capturing, sharing & recording in SwiftUI/UIKit. Retrieved April 2025, from https://medium.com/@Lakshmnaidu/securing-app-from-screen-capturing-sharing-recording-apps-swiftui-uikit-26057810c90d utkarshkore.medium.com+4medium.com+4support.google.com+4

Wickramathilaka, S., Grundy, J., Madampe, K., & Haggag, O. (2025). Adaptive and accessible user interfaces for seniors through model-driven engineering. arXiv. https://arxiv.org/abs/2502.18828

Zeexan. (2025). detect_fake_location: A Flutter plugin for detecting mock location. Pub.dev. Retrieved June 2025, from https://pub.dev/packages/detect_fake_location

Zeng, G., Chen, S., Mu, B., Shi, G., & Wu, J. (2023, May). Cpnp: Consistent pose estimator for perspective-n-point problem with bias elimination. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1940-1946). IEEE.

CHAPTER 9

A Historical Overview of Language Model Development: From Statistical Models to Transformers

Ahmet Toprak[1]
Feyzanur Sağlam Toprak[2]

1. Introduction

Language modeling lies at the heart of natural language processing (NLP), providing the foundation for machines to understand, generate, and interact with human language. From early rule-based systems to today's state-of-the-art deep learning architectures, the development of language models has marked a profound transformation in how computational systems process textual information. A language model estimates the probability distribution over sequences of words, enabling tasks such as text completion, translation, question answering, and dialogue generation.

The evolution of language models can be viewed as a reflection of broader advancements in both linguistic theory and machine learning techniques. Initial models were grounded in statistical methods, relying on the frequency and co-occurrence of words within corpora. These approaches, while simple and interpretable, were limited by their inability to capture long-range dependencies and contextual nuance.

[1] Ahmet Toprak, Department of Computer Engineering, Istanbul Ticaret University, Istanbul/Türkiye, 0000-0001-7046-8512, ce.ahmet.toprak@gmail.com
[2] Feyzanur Sağlam Toprak, Türkiye Finans Participation Bank, Istanbul/ Türkiye, 0000-0000-0000-0000, feyza-saglam@hotmail.com

As the demand for more intelligent language systems grew, particularly with the emergence of large-scale digital text and increased computational power—researchers began to explore more sophisticated approaches. The introduction of Hidden Markov Models (HMMs) and n-gram techniques provided a probabilistic framework for sequence modeling, but they remained constrained by fixed context windows and issues related to sparsity.

The advent of neural networks in NLP marked a paradigm shift. Starting with feed-forward neural language models, the field quickly progressed to recurrent neural networks (RNNs) [1, 2], which introduced the capacity to model sequential dependencies adaptively and dynamically. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) further addressed the vanishing gradient problem, enabling deeper and more meaningful temporal representations of language. Yet even these architectures encountered limitations in parallelization and long-term context retention. The breakthrough came with the introduction of the Transformer architecture, which replaced recurrence with self-attention mechanisms, allowing for greater scalability and context awareness. Transformers now power the most advanced language models, including BERT, GPT, T5, and their successors, which have redefined the boundaries of what is possible in language understanding and generation.

This paper presents a comprehensive review of this trajectory—from statistical foundations to the era of massive pretrained Transformer models. In doing so, we examine how each phase in this progression addressed prior limitations, contributed novel capabilities, and influenced downstream NLP applications such as machine translation, summarization, and conversational AI. By tracing this history, we aim to contextualize current innovations and highlight future directions in language modeling research.

2. Early Statistical Language Models

Before the advent of neural architecture and deep learning, statistical language models were the dominant method for enabling machines to process and generate human language. These models operate by estimating the likelihood of a given word or sequence of words based on their observed frequencies in large text corpora. Despite their limitations, early statistical models played a critical role in shaping the field of natural language processing (NLP) and served as the foundation upon which more complex systems were later built.

2.1.N-Gram-Based Models

Among the earliest and most influential statistical language models are n-gram models. In essence, an n-gram model predicts the next word in a sequence by looking at the previous one or more words. The number “n” in n-gram refers to the length of the word sequence being considered. For example, a unigram model looks at single words independently, a bigram model considers pairs of consecutive words, and a trigram model examines sequences of three words.

These models operate under the simplifying assumption that a word’s occurrence depends only on a limited history of preceding words. This makes them computationally tractable and easy to implement. Training n-gram models involves collecting large volumes of text and counting how often certain sequences of words occur. The probability of new sequences is then inferred based on these frequency counts.

The main strength of n-gram models lies in their simplicity and speed. They are effective at capturing local patterns and are particularly useful for applications such as spell correction, predictive typing, and speech recognition. Moreover, when combined with smoothing techniques—used to assign small probabilities to previously unseen sequences—they can generalize beyond strictly observed data.

However, n-gram models suffer from several significant limitations. First, their reliance on fixed-length context windows means they are unable to model long-range dependencies. For instance, understanding that the subject and verb in a sentence agree in number often requires analyzing words that are several positions apart—something n-gram models cannot do effectively. Second, the larger the value of "n," the more data is required to produce reliable estimates. This leads to problems with data sparsity, especially when dealing with rare words or unusual constructions. Third, these models treat words as discrete units and do not capture semantic relationships between them, meaning they cannot generalize well to synonyms or related phrases.

Despite these drawbacks, n-gram models were widely used throughout the 1990s and early 2000s. Their efficiency and straightforward implementation made them a default choice for many practical NLP systems prior to the emergence of data-driven machine learning approaches.

2.2. Hidden Markov Models

To overcome some of the shortcomings of n-gram models, researchers turned to probabilistic models that could incorporate hidden structure—most notably, the Hidden Markov Model (HMM). HMMs are a class of generative models that assume a sequence of observable events (such as words in a sentence) is generated by a sequence of hidden states (such as syntactic or semantic categories). While the observed data can be directly measured, the underlying state sequence is inferred from the data.

In language processing tasks, HMMs proved especially useful for applications such as part-of-speech tagging, named entity recognition, and speech recognition. By modeling the underlying grammatical or phonetic structure of language as a series of hidden states, HMMs can capture more nuanced patterns in sequences than simple n-gram models.

One of the key advantages of HMMs is their ability to manage uncertainty and make inferences based on incomplete or noisy data. The probabilistic nature of the model allows it to weigh different possible interpretations and select the most likely one. This is particularly useful in real-world scenarios where input data may be ambiguous or prone to errors, such as in speech recognition or machine translation.

Nonetheless, HMMs also have limitations. They assume that the probability of a state depends only on the previous state (the Markov assumption) and that the observed word depends solely on the current state. These assumptions limit the model's ability to capture complex dependencies and interactions across longer spans of text. Moreover, like n-gram models, HMMs represent words as isolated symbols, without any consideration for meaning or similarity. This makes them less adaptable to novel or out-of-vocabulary words and restricts their performance on tasks requiring deep contextual understanding.

Despite these constraints, HMMs represented a significant advancement over simpler statistical models. They introduced the idea of modeling latent structure in language concept that would later become central in neural network-based approaches. HMMs remained the state-of-the-art in many NLP tasks until the early 2010s, when they were gradually supplanted by neural architectures that offered greater expressive power and flexibility.

3. Neural Language Models

The transition from statistical to neural language models marked a transformative phase in the development of natural language processing. While statistical models such as n-grams and Hidden Markov Models provided useful approximations of linguistic patterns, they were fundamentally limited by their reliance on fixed context windows and discrete word representations. Neural language models addressed these limitations by introducing continuous representations and dynamic context modeling, thereby enabling more expressive and flexible systems.

3.1.The Emergence of Neural Probabilistic Language Models

The first breakthrough in neural language modeling came with the neural probabilistic language model (NPLM) proposed by Bengio et al. in 2003. This model introduced the concept of representing words as dense vectors (word embeddings) in a continuous space, where semantically similar words are located closer together. Instead of estimating the probability of a word sequence using counts, NPLM used a feed-forward neural network to learn word dependencies and context representations in a data-driven manner.

This innovation addressed several of the key limitations of n-gram models. First, the use of embeddings enabled generalization to unseen word combinations. Second, neural networks could learn non-linear interactions between words, capturing more complex linguistic relationships. However, feed-forward models were still restricted by their fixed-size context window, and they required retraining to accommodate new sequences or vocabulary.

3.2.Recurrent Neural Networks (RNNs)

To overcome the limitations of fixed context, researchers introduced Recurrent Neural Networks (RNNs)—a class of neural networks designed to model sequential data by maintaining a dynamic internal state that evolves over time. In RNNs, information from previous inputs is preserved and passed forward, allowing the model to theoretically capture dependencies across arbitrarily long sequences [3].

RNNs [4, 5] significantly improved performance in tasks such as language modeling, text generation, and machine translation. However, they suffered from practical challenges, most notably the vanishing and exploding gradient problems—which made it difficult for the model to retain information over long sequences. This issue hindered their ability to learn long-term dependencies, a critical requirement for natural language understanding.

3.3.Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)

To address the shortcomings of standard RNNs, more advanced architectures were developed. The most influential of these is the Long Short-Term Memory (LSTM) network, introduced by Hochreiter and Schmidhuber in 1997 [6, 7]. LSTMs incorporate a memory cell and a set of gates—input, forget, and output—that regulate the flow of information through the network. This structure enables LSTMs to retain important information over long sequences while filtering out irrelevant details.

LSTMs [8, 9] became the backbone of many state-of-the-art NLP systems in the 2010s, powering breakthroughs in speech recognition, sentiment analysis, and question answering. A simpler alternative to LSTM, known as the Gated Recurrent Unit (GRU), was later introduced and offered comparable performance with fewer parameters and more streamlined architecture.

These gated recurrent networks [10] represented a significant step forward in language modeling. They allowed models to maintain richer contextual awareness and adapt to varied sentence structures. However, despite their strengths, they still operated sequentially, making them inefficient to train and difficult to parallelize, particularly on large datasets.

By the mid-2010s, the limitations of recurrence-based models created a bottleneck in scaling up language models. This paved the way for the next major leap: the Transformer architecture, which would redefine the landscape of NLP and usher in a new era of performance and generalization.

4. The Rise of Transformer-Based Models

The introduction of Transformer architecture marked a paradigm shift in the field of natural language processing. Presented by Vaswani et al. in the seminal 2017 paper “Attention Is All You Need”, the Transformer model replaced the sequential nature of recurrent neural networks with a fully attention-based mechanism. This innovation significantly enhanced the efficiency, scalability, and performance of language models across a wide range of NLP tasks.

4.1.Architectural Innovation

Unlike RNNs and LSTMs, which process input sequences token by token, Transformers process entire sequences in parallel using self-attention mechanisms. The self-attention layer allows each word in a sentence to attend to every other word, enabling the model to learn global dependencies without regard to word order [11, 12, 13].

This parallelism offers two major benefits: first, it allows for significant speedup during training, making it possible to scale models to massive datasets. Second, it improves the model's ability to capture long-range dependencies, which are often crucial for understanding complex language constructs.

The core components of the Transformer include:

- Multi-head self-attention: Allows the model to focus on different parts of the sequence simultaneously.
- Positional encoding: Adds information about word order, compensating for the loss of sequential structure.
- Feed-forward layers and layer normalization: Improve learning dynamics and representation power.

These design choices make Transformers not only more expressive than RNNs but also better suited for pretraining on large corpora, a key characteristic of modern language models.

4.2.Emergence of Pretrained Transformer Models

Building on the Transformer architecture, a series of pretrained language models emerged, demonstrating unprecedented performance on NLP benchmarks. These models follow a two-phase training strategy [14, 15]:

1. Pretraining: The model learns general language representations by predicting missing words (masked language modeling), next sentences, or future tokens over massive datasets.
2. Fine-tuning: The pretrained model is adapted to specific tasks such as classification, summarization, or translation using task-specific labeled data.

Notable Transformer-based models include:

- BERT (Bidirectional Encoder Representations from Transformers): Introduced by Devlin et al. in 2018, BERT utilizes masked language modeling and next sentence prediction. Its bidirectional attention mechanism allows it to capture context from both sides of a target word, making it highly effective for understanding tasks like question answering and natural language inference [16].
- GPT (Generative Pretrained Transformer): Developed by OpenAI, GPT uses unidirectional attention and is trained via causal language modeling, predicting the next word in a sequence. Its generative nature makes it ideal for text completion, dialogue generation, and creative writing tasks. The successive versions (GPT-2, GPT-3, GPT-4) have demonstrated increasing scale and capabilities [17].

- T5 (Text-to-Text Transfer Transformer): Proposed by Google Research, T5 formulates every NLP task as a text-to-text problem, unifying various tasks under a single framework and achieving high versatility.
- XLNet, RoBERTa, ALBERT, and others: These models introduced architectural refinements or training optimizations to further enhance performance, such as improved pretraining objectives, parameter sharing, and longer training times [18].

4.3.Impact and Capabilities

Transformer-based models have revolutionized NLP by achieving state-of-the-art results on virtually every benchmark, including GLUE, SuperGLUE, SQuAD, and more. Their impact extends beyond academic research into real-world applications such as:

- Intelligent virtual assistants (e.g., Siri, Alexa, ChatGPT)
- Legal and biomedical document analysis
- Automated customer support
- Code generation and reasoning
- Real-time translation and summarization

Moreover, the concept of foundation models—large pretrained models capable of generalizing across diverse tasks—has become a defining trend in AI research.

While these models offer unprecedented capabilities, they also introduce challenges related to computational cost, environmental impact, interpretability, and potential misuse. Addressing these concerns remains an active area of research as the community seeks to build safer, fairer, and more efficient AI systems.

5. Applications and Implications for Nlp Tasks

The evolution of language models—from statistical methods to deep learning-based architecture has significantly expanded the range and effectiveness of natural language processing (NLP) applications. As models have become more context-aware, scalable, and semantically robust, their impact on downstream NLP tasks has grown exponentially. This section explores how language models have transformed key NLP domains such as machine translation, text generation, and automatic summarization, while also addressing the broader implications of these advancements.

5.1.Machine Translation

Machine translation is one of the most prominent and impactful applications of language modeling. Early systems were built on rule-based or phrase-based statistical approaches, which required extensive linguistic knowledge and hand-crafted alignment strategies. The introduction of neural machine translation (NMT) systems—powered initially by RNNs and later by Transformers—marked a major improvement in translation quality, fluency, and adaptability.

Transformer-based models such as Transformer-Base, BERT, and mBART have significantly improved cross-lingual understanding by learning multilingual representations. These models capture syntactic and semantic regularities across languages, allowing for zero-shot or few-shot translation capabilities. Moreover, large-scale pretrained models like mT5 and XLM-R have extended translation capabilities to low-resource languages, democratizing access to multilingual AI technologies.

5.2.Text Generation

Text generation tasks involve producing coherent and contextually appropriate text given an input prompt or initial sequence. Traditional statistical models struggled with fluency and creativity, often producing repetitive or ungrammatical output. Recurrent models such as LSTMs offered better coherence, but they were still limited in long-range planning and topic consistency.

The advent of models like GPT-2 [19], GPT-3 [20], and beyond has redefined expectations for text generation. These models can produce human-like text across diverse genres—stories, articles, dialogues, code, and even poetry. Their ability to maintain context over extended passages and generate responses that reflect complex reasoning makes them suitable for applications such as:

- Conversational agents and chatbots
- Creative writing and content drafting
- Automated report generation
- Email and message completion

Despite their capabilities, concerns remain regarding factual accuracy, bias propagation, and ethical use—particularly when these systems are deployed at scale or in sensitive contexts.

5.3.Text Summarization

Text summarization aims to condense large volumes of text into shorter, informative versions without losing essential meaning. Early methods relied on extractive techniques—selecting and reordering existing sentences from the input. While effective in capturing key phrases, such approaches often lacked coherence and failed to capture implicit meaning.

With the emergence of pretrained encoder-decoder models like T5 [21], BART [22], and PEGASUS [23], abstractive summarization has become feasible and increasingly accurate. These models generate novel summaries that rephrase and synthesize information, closely mimicking how humans summarize content.

Applications include:

- Summarizing news articles or academic papers
- Creating executive summaries for reports
- Email thread summarization
- Legal or medical document abstraction

Transformer-based models have made summarization more scalable and adaptable across domains. However, they also introduce the risk of generating hallucinated content—text that is fluent but factually incorrect, which remains a key area for further research.

5.4.Broader Implications

The widespread deployment of language models in NLP has implications that extend beyond task performance. On one hand, these models are enabling new levels of accessibility, productivity, and automation. On the other hand, they raise questions about computational cost, environmental sustainability, data privacy, and social impact.

Issues such as model bias, cultural representation, and content moderation are increasingly relevant as models interact with users across different languages, backgrounds, and intentions. Addressing these challenges requires not only technical innovation but also interdisciplinary collaboration, regulatory insight, and ethical foresight.

6. Conclusion

The development of language models has undergone a remarkable transformation over the past several decades, evolving from simple statistical approximations to complex neural architectures capable of understanding and generating human-like language. Each stage in this evolution—unigram and bigram models, Hidden Markov Models, recurrent neural networks, and finally Transformer-based systems—has contributed to overcoming the limitations of its predecessors and expanding the scope of what is achievable in natural language processing.

Early statistical models provided a foundation for understanding linguistic probabilities, offering computationally efficient solutions despite their restricted context awareness. With the rise of neural networks, particularly RNNs and LSTMs, models gained the ability to capture sequential dependencies and dynamic structure in language. However, these models were still constrained by issues such as training inefficiency and difficulty in handling long-range context.

The advent of the Transformer architecture has revolutionized the field, enabling massive scale, parallel processing, and nuanced understanding of global linguistic relationships. Transformer-based models such as BERT, GPT, T5, and their derivatives have set new performance benchmarks across nearly all major NLP tasks, making them the standard for modern language technologies. These models not only achieve higher accuracy but also open the door for new applications in multilingual processing, conversational AI, summarization, and cross-domain adaptation.

Despite these advancements, the journey is far from over. Several key directions are likely to shape the next phase of language modeling:

- Efficiency and Sustainability: As models grow in size and complexity, so do their environmental and financial costs. Future research must prioritize model compression, energy-efficient training, and lightweight deployment strategies, particularly for use in low-resource and mobile environments.
- Explainability and Transparency: With the increasing adoption of language models in high-stakes domains such as healthcare, law, and education, there is a pressing need to make them more interpretable. Understanding how and why a model generates a particular output is crucial for building trust and accountability.
- Fairness and Ethical AI: Large language models often reflect and amplify societal biases present in their training data. Addressing these issues requires robust fairness auditing, bias mitigation techniques, and inclusive dataset curation to ensure equitable outcomes for all users.
- Multimodality and Integration: The future of language modeling is likely to include deeper integration with other modalities—such as vision, audio, and structured data—paving the way for more comprehensive AI systems capable of holistic perception and reasoning.
- Interactive and Continual Learning: Instead of static pretraining, emerging paradigms may emphasize models that learn continuously from user interaction, adapt to specific domains, and evolve with minimal supervision.

In conclusion, language modeling stands as one of the most dynamic and influential areas in artificial intelligence. By reflecting on its historical trajectory, we gain valuable insight into both the technical milestones and the conceptual shifts that have shaped the field. As we move forward, interdisciplinary collaboration, ethical stewardship, and responsible innovation will be essential to harness the full potential of language models in ways that are beneficial, equitable, and aligned with human values.

7. References

[1] Y. Chen and J. Li, "Recurrent Neural Networks algorithms and applications," 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Zhuhai, China, 2021, pp. 38-43, doi: 10.1109/ICBASE53849.2021.00015.

[2] M. Kaur and A. Mohta, "A Review of Deep Learning with Recurrent Neural Network," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 460-465, doi: 10.1109/ICSSIT46314.2019.8987837.

[3] N. M. Rezk, M. Purnaprajna, T. Nordström and Z. Ul-Abdin, "Recurrent Neural Networks: An Embedded Computing Perspective," in IEEE Access, vol. 8, pp. 57967-57996, 2020, doi: 10.1109/ACCESS.2020.2982416.

[4] S. Sivamohan, S. S. Sridhar and S. Krishnaveni, "An Effective Recurrent Neural Network (RNN) based Intrusion Detection via Bi-directional Long Short-Term Memory," 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498552.

[5] E. Diao, J. Ding and V. Tarokh, "Restricted Recurrent Neural Networks," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 56-63, doi: 10.1109/BigData47090.2019.9006257.

[6] S. Yang, X. Yu and Y. Zhou, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI), Shanghai, China, 2020, pp. 98-101, doi: 10.1109/IWECAI50956.2020.00027.

[7] R. Fu, Z. Zhang and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China, 2016, pp. 324-328, doi: 10.1109/YAC.2016.7804912.

[8] Y. Liu, "Stock Prediction Using LSTM and GRU," 2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 2022, pp. 206-211, doi: 10.1109/ICDSBA57203.2022.00054.

[9] E. Tuna and A. Soysal, "LSTM and GRU based Traffic Prediction Using Live Network Data," 2021 29th Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9478011.

[10] G. Goui, A. Zrelli and N. Benletaief, "A Comparative Study of LSTM/GRU Models for Energy Long-Term Forecasting in IoT Networks," 2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS), Wuxi, China, 2023, pp. 60-64, doi: 10.1109/ICIS57766.2023.10210257.

[11] J. Wu, X. Huang, J. Liu, Y. Huo, G. Yuan and R. Zhang, "NLP Research Based on Transformer Model," 2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom), Xiangtan, Hunan, China, 2023, pp. 343-348, doi: 10.1109/CSCloud-EdgeCom58631.2023.00065.

[12] R. Kora and A. Mohammed, "A Comprehensive Review on Transformers Models For Text Classification," 2023 International Mobile, Intelligent, and

Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 2023, pp. 1-7, doi: 10.1109/MIUCC58832.2023.10278387.

[13] F. Passos et al., "Machine Learning Approaches for Transformer Modeling," 2022 18th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), Villasimius, Italy, 2022, pp. 1-4, doi: 10.1109/SMACD55068.2022.9816303.

[14] R. C. Dugan, "A perspective on transformer modeling for distribution system analysis," 2003 IEEE Power Engineering Society General Meeting (IEEE Cat. No.03CH37491), Toronto, ON, Canada, 2003, pp. 114-119 Vol. 1, doi: 10.1109/PES.2003.1267146.

[15] A. Rawat and S. Singh Samant, "Comparative Analysis of Transformer based Models for Question Answering," 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2022, pp. 1-6, doi: 10.1109/CISCT55310.2022.10046525.

[16] Y. O. Sharrab, H. Attar, M. A. H. Eljinini, Y. Al-Omary and W. E. Al-Momani, "Advancements in Speech Recognition: A Systematic Review of Deep Learning Transformer Models, Trends, Innovations, and Future Directions," in IEEE Access, vol. 13, pp. 46925-46940, 2025, doi: 10.1109/ACCESS.2025.3550855.

[17] Q. Luo, W. Zeng, M. Chen, G. Peng, X. Yuan and Q. Yin, "Self-Attention and Transformers: Driving the Evolution of Large Language Models," 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), Qingdao, China, 2023, pp. 401-405, doi: 10.1109/

[18] A. Gillioz, J. Casas, E. Mugellini and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," 2020 15th Conference on Computer Science and Information Systems (FedCSIS), Sofia, Bulgaria, 2020, pp. 179-183, doi: 10.15439/2020F20.

[19] G. Yenduri et al., "GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," in IEEE Access, vol. 12, pp. 54608-54649, 2024, doi: 10.1109/ACCESS.2024.3389497.

[20] B. Chen et al., "On the Use of GPT-4 for Creating Goal Models: An Exploratory Study," 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, 2023, pp. 262-271, doi: 10.1109/REW57809.2023.00052.

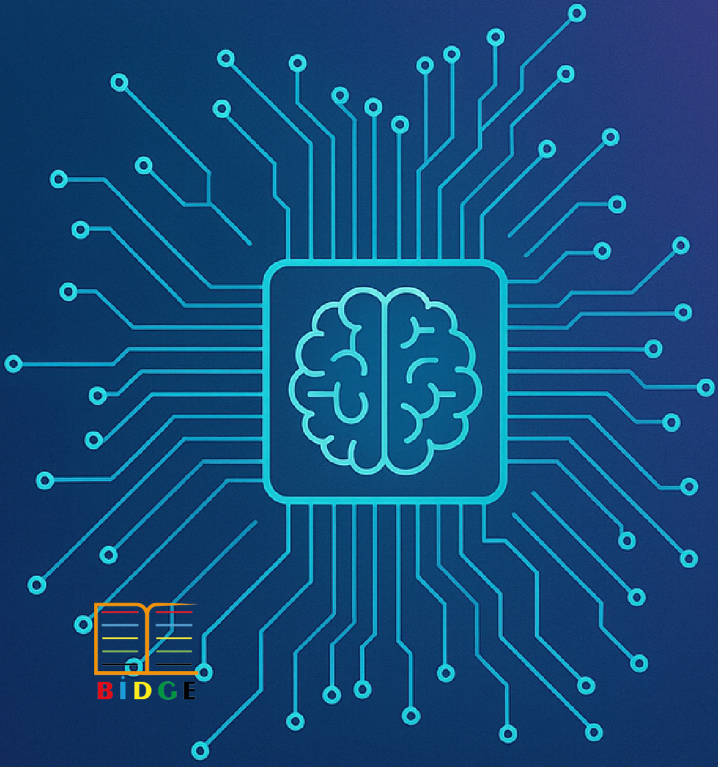
[21] S. S. Kalakonda, S. Maheshwari and R. K. Sarvadevabhatla, "Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation," 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 2023, pp. 31-36, doi: 10.1109/ICME55011.2023.00014.

[22] P. J. Giabbanelli, "GPT-Based Models Meet Simulation: How to Efficiently use Large-Scale Pre-Trained Language Models Across Simulation Tasks," 2023 Winter Simulation Conference (WSC), San Antonio, TX, USA, 2023, pp. 2920-2931, doi: 10.1109/WSC60868.2023.10408017.

[23] A. Gillioz, J. Casas, E. Mugellini and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," 2020 15th Conference on Computer Science and Information Systems (FedCSIS), Sofia, Bulgaria, 2020, pp. 179-183, doi: 10.15439/2020F20.

AI-DRIVEN DIGITAL ECOSYSTEMS

FROM GENETICS TO IOT FOR A SECURE AND SUSTAINABLE FUTURE



BIDGE