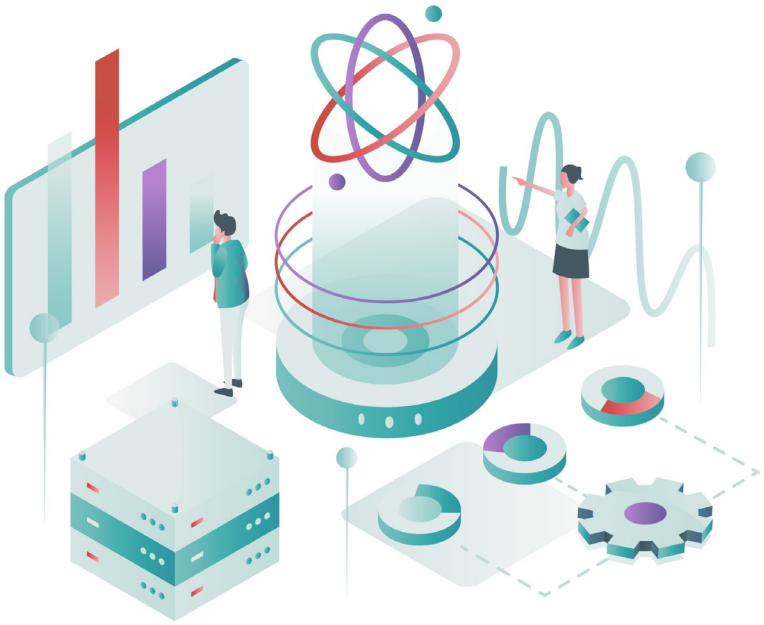
CURRENT STUDIES IN DATA SCIENCE AND ANALYTICS

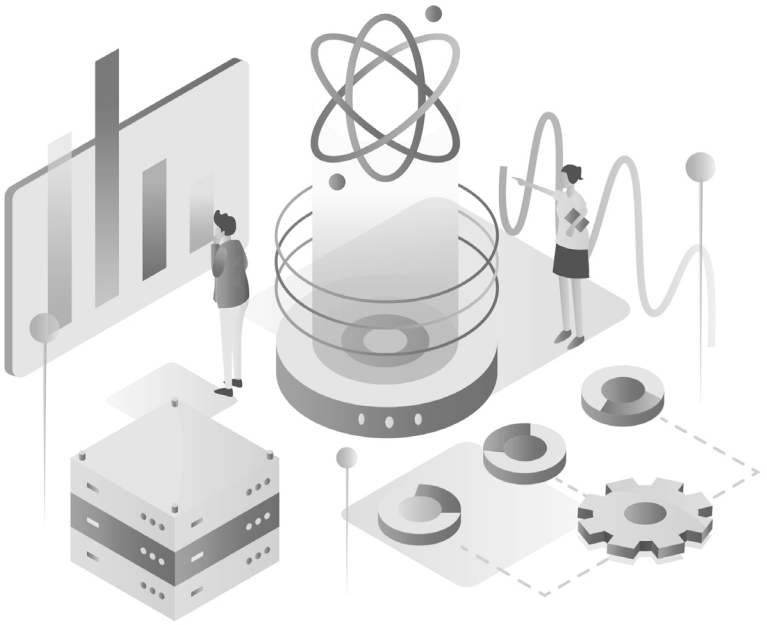


Editors

M. Hanefi CALP
Resul BÜTÜNER

ISRES Publishing

CURRENT STUDIES IN DATA SCIENCE AND ANALYTICS



Editors

M. Hanefi CALP
Resul BÜTÜNER

ISRES Publishing

Current Studies in Data Science and Analytics

Edited by

Assoc. Prof. Dr. M. Hanefi CALP
Ankara Hacı Bayram Veli University, Faculty of Economics and Administrative Sciences, Management Information Systems, Ankara, Türkiye
Email: hanefi.calp@hbv.edu.tr

Engineer, MSc. Resul BÜTÜNER
Ministry of National Education,
Directorate General for Innovation and Educational Technologies, Ankara, Türkiye
Email : resul.butuner@eba.gov.tr

Language Editor

Prof. Dr. Kadriye Dilek BACANAK
Gazi University, Faculty of Education, Foreign Languages Education Department, English Language Teaching Division, Ankara, Türkiye
Email: kadriyedilek@gmail.com

ISRES Publishing

Current Studies in Data Science and Analytics

Editors

M. Hanefi CALP

Resul BÜTÜNER

This book was typeset in 10/12 pt. Times New Roman, Italic, Bold and Bold Italic.

Current Studies in Data Science and Analytics

Published by ISRES Publishing, International Society for Research in Education and Science (ISRES). Includes bibliographical references and index.

ISBN

978-625-6959-43-9

Date of Issue

December, 2024

Contact

Askan Mah. Akinbey Sok. No: 5/A Meram/Konya/Türkiye
isresoffice@gmail.com
www.isres.org

About the Book

In an era where the innovation and progress of data, the field of Data Science has emerged as a science, unlocking transformative insights across a multitude of domains. With data becoming very important in decision-making, organizations, educators, and researchers are increasingly relying on advanced analytical techniques to navigate complexities and drive meaningful outcomes. This book, Current Studies in Data Science and Analytics, brings together a collection of contemporary research and practical applications, offering readers a comprehensive look at the evolving landscape of Data Science. From data analytics to data visualization, geographic information systems to cybersecurity, education, and gaming, underscores the versatility and relevance of data-driven methodologies.

This book was prepared from the selected academic research and review studies invited by the editors. It was focused on the issues based on "Current Studies in Data Science and Analytics" under the leadership of The International Society for Research in Education and Science (ISRES) Publishing which is a Recognized International Publishing. In addition, all submissions were reviewed by at least two international referees and composed of 10 chapters (15 authors) selected by the editors. The purpose of the book is to provide readers with the opportunity to receive a scholarly refereed publication in the field of data science and analytics. Finally, we hope the book will present curiosity in this field, the book will be useful for new scientists, science readers, and anyone who intends to learn about the mystery of science.

December, 2024

Assoc. Prof. Dr. M. Hanefi CALP
Ankara Hacı Bayram Veli University

Engineer, MSc. Resul BÜTÜNER
Ministry of National Education

Table of Contents

Chapter 1 — 1-19
Data-Driven Analytical Techniques in Geographic Information Systems
Murat KILINÇ

Chapter2 — 20-27
The New Role of Teachers in the Age of Web 3.0: Personalized Learning Environments through Learning Analytics
Hüseyin ULUKUZ

Chapter3 — 28-37
Pioneering Data-Driven Decisions: The Future of Predictive Modeling in Data Science
Kartal DERİN

Chapter 4 — 38-56
An Overview of Social Network Analysis: Metrics, Tools and Applications
Akça Okan YÜKSEL

Chapter 5 — 57-85
Data Science Applications in Games
Murat ATASOY, Adil YILDIZ, Lokman ŞILBIR, Ekrem BAHÇEKAPILI

Chapter 6 — 86-96
Artificial Intelligence-Powered Data Analytics against Botnet Attacks: Threat Detection and Ethical Considerations
Ramazan KOCAOĞLU

Chapter 7 — 97-108
Cyber Threat Analytics in Data Science: Intrusion Detection And Prevention Systems
Özgür TONKAL

Chapter 8 — 109-123
Navigating the Data Science Landscape: Essential Competencies
Mehmet KOKOÇ

Chapter 9 — 124-143
Analysis of IoT Security Datasets
Erdal ÖZDOĞAN, Onur CERAN

Chapter 10 — 144-155
The Importance of Dashboard in Data Analysis: An Application Example
M. Hanefi CALP, Resul BÜTÜNER

Citation

Calp, M. H., & Bütüner, R. (Eds.). (2024). *Current Studies in Data Science and Analytics.* ISRES Publishing.

Managing Editors

Assoc. Prof. Dr. M. Hanefi CALP received Ph.D. degree from the department of Management Information Systems at Gazi University, one of the most prestigious universities in Türkiye. He works as an Associate Professor in the Department of Management Information Systems of the Faculty of Economics & Administrative Sciences of the Ankara Hacı Bayram Veli University. His research interest includes Management Information Systems, Digital Transformation, Artificial Neural Networks, Expert Systems, Fuzzy Logic, Risk Management, Risk Analysis, Human-Computer Interaction, Technology Management, Knowledge Management, and Project Management.
E-mail: hanefi.calp@hbv.edu.tr , **ORCID:** 0000-0001-7991-438X

Resul BÜTÜNER is an information technology teacher at the Ministry of National Education in Ankara, Türkiye. He has a master's degree in Computer Engineering from Necmettin Erbakan University. He is currently working on a book in the field of artificial intelligence, robotic coding, data mining, and augmented reality applications. He is an instructor in the field of Robotic coding within TUBITAK. He continues to write a book in the field of robotic coding at the Ministry of National Education. He worked as a coordinator in projects related to student education. .
E-mail: resul.butuner@eba.gov.tr , **ORCID:**0000-0002-9778-2349

In This Book

Chapter 1,

This study explores how Geographic Information Systems (GIS) can be enhanced through the integration of big data, data science, and machine learning techniques. It highlights the role of data-driven approaches in overcoming the limitations of traditional GIS methods.This provides new dimensions to spatial data analysis. The study focuses on the utilization of machine learning and deep learning techniques for processing and analyzing big data obtained from various sources such as satellite imagery, sensors, and social media. Application areas such as urban planning, disaster management, environmental monitoring, and transportation analysis are discussed as examples. Additionally, the study examines the advantages of integrating augmented reality (AR), real-time data analytics, and cloud-based solutions to GIS. These technologies are shown to have significant potential in areas such as city planning, traffic management, and monitoring environmental changes. The importance of data visualization tools and techniques in facilitating the interpretation of spatial data and supporting decision-making processes is emphasized. Finally, the study addresses existing challenges, including data quality, integration issues, and high computational costs, while discussing future trends such as AI-powered models and cloud-based solutions.

Chapter 2,

In the age of Web 3.0, the role of teachers is undergoing a significant transformation as personalized learning environments become more prevalent through the use of learning analytics. With the advent of advanced digital tools and platforms, educators are no longer just content providers. Instead, they are becoming facilitators who guide students on individualized learning paths and moderators who manage dynamic learning environments to foster collaboration, engagement, and effective communication. Learning analytics enables teachers to harness data-driven insights to tailor educational experiences that align with each student's unique strengths, weaknesses, and interests. This shift enables a more targeted approach to teaching, where educators can identify and address learning gaps promptly, fostering a more inclusive and effective learning atmosphere. By leveraging learning analytics, teachers can interact effectively with their students, creating a dynamic educational landscape that emphasizes collaboration and lifelong learning in an increasingly digital world.

Chapter 3,

Predictive modeling stands at the forefront of data-driven decision-making, transforming industries by leveraging historical data through advanced machine learning and deep learning techniques. From early disease detection in healthcare from supply chain optimization to personalized education, these models are revolutionizing operations

with remarkable precision. As technologies like quantum computing, federated learning, and edge AI enhance their speed, scalability, and accuracy, predictive analytics continues to unlock unprecedented opportunities across sectors. However, addressing ethical concerns such as bias, transparency, and accountability remains critical for responsible innovation. By navigating these challenges, predictive modeling promises a future where data-driven insights empower ethical and impactful decision-making on a global scale.

Chapter 4,

The document provides a comprehensive overview of Social Network Analysis (SNA), emphasizing its foundational concepts, metrics, tools, and applications. SNA examines relationships within networks of individuals, groups, or organizations, utilizing methods like degree, betweenness, closeness, eigenvector, and pagerank centralities, local clustering coefficient, density, size, average degree, centralization to analyze nodes' roles and connectivity. Historical development highlights its evolution from formal sociology to a distinct interdisciplinary field supported by software tools like Gephi, UCINET, NodeXL, PAJEK, R Programming packages and NetworkX. Applications span education, healthcare, and business, with growing relevance in handling large, complex data. Current trends focus on artificial intelligence, privacy, and advanced network structures, underscoring SNA's expanding significance in academic and commercial contexts. Finally, examples demonstrating the calculation of metrics within a social network are provided.

Chapter 5

This chapter examines the profound impact of data science on the digital gaming industry, showcasing its ability to drive innovation and reshape key aspects of game development, player experience, and business strategies. Data science serves as a transformative force, enabling the optimization of game mechanics, personalization of player interactions, and creation of dynamic, adaptive gaming environments. By leveraging advanced analytics, artificial intelligence, and machine learning, developers gain deep insights into player behavior, allowing for more engaging and balanced gameplay. Emerging technologies like blockchain, edge computing, and emotional design further enhance the industry, offering secure asset management, low-latency gaming experiences, and emotionally intelligent interactions. These advancements not only redefine the technical and creative possibilities of games but also pave the way for new monetization models and immersive, player-centered ecosystems. As data-driven practices become increasingly integral to gaming, ethical considerations and privacy concerns also come to the forefront, requiring transparent and responsible approaches. This chapter explores these multifaceted dimensions, illustrating how data science continues to revolutionize the gaming world.

Chapter 6,

In this chapter, how AI-supported data analytics methods can be used against botnet attacks, the effectiveness of these technologies in security, the technical and ethical challenges encountered, and future development areas will be evaluated in detail. At the end of the chapter, in the light of the solutions offered by artificial intelligence against botnet attacks, recommendations will be presented for future directions and the development of healthier solutions in this field.

Chapter 7,

Data science provides a effective analysis tool for identifying threat patterns and predicting potential attacks based on historical data of cybersecurity incidents. This section will focus on how data science approaches are integrated into critical security structures such as intrusion detection systems (IDS) and intrusion prevention systems (IPS) used to detect and prevent cyber threats. In this context, important techniques such as machine learning, big data analytics, and anomaly detection methods will be detailed. The aim of the section is to understand the critical role of data science techniques in cybersecurity and to demonstrate how these techniques are applied in practice. At the same time, it is aimed to inform readers about the current state of cybersecurity analytics and to provide a glimpse into where this field may evolve in the future.

Chapter 8,

This chapter examines the fundamental competencies required in the field of data science, providing a detailed analysis of its historical development, current practices, and future potential. It emphasizes the interdisciplinary nature of data science, showcasing its reliance on the integration of statistics, computer science, and mathematics to solve complex, real-world problems. The chapter highlights the importance of a balanced skill set, combining technical expertise, analytical abilities, and soft skills, to meet the increasing demand for data scientists across various industries. The chapter outlines key technical competencies, including programming, data analysis, machine learning, big data technologies, and data visualization. In addition, it underscores the value of soft skills such as effective communication, teamwork, adaptability, and ethical awareness, which are essential for navigating dynamic and collaborative work environments. The role of educational and training programs in fostering these skills is also explored, with a focus on aligning academic curricula with industry expectations.The chapter further explores the transformative impact of emerging technologies, including artificial intelligence, cloud computing, and quantum computing, on the future of data science. It also considers critical issues such as data privacy, algorithmic fairness, and ethical responsibility. By offering a structured perspective, this chapter serves as a practical resource for professionals aiming to advance their careers in data science and for

educators seeking to design comprehensive training programs that address both technical and ethical challenges.

Chapter 9,

In this book chapter, the focus is on analyzing and evaluating datasets commonly used in Intrusion Detection Systems for IoT security. Given the complex networks formed by interconnected devices, the chapter emphasizes the critical role of effective data analysis and attack detection in enhancing IoT system security. Through a multifaceted examination, the study identifies key features and pre-processing requirements of datasets that contribute to improved model performance. Additionally, it investigates the diversity of attack types within these datasets, aiding researchers in selecting the most appropriate datasets for specific scenarios. By conducting comparative analyses, the chapter aims to highlight the most suitable datasets for IoT IDS systems, contributing valuable insights to the field of cybersecurity.

Chapter 10,

In this book chapter, the study aimed to find answers to the extent to which the COVID-19 vaccines produced and used against the virus in the COVID-19 pandemic are related to basic headings such as supply and vaccine inequality. In this context, the data analysis process carried out in the study was supported by dashboards prepared using the Power BI program. Thus, the data will be analyzed more effectively and efficiently in a shorter time. These indicators will also make it easier for institutions, organizations, or individuals to make accurate, efficient, and effective comments and thus make some strategic decisions. In order to achieve the objectives of the study to the maximum extent, the these questions were sought to be answered: Are states able to deliver vaccines to their citizens equally during the pandemic period? If not, what are the reasons for this? What are the effects of vaccine inequality? How can vaccine inequality be resolved? What are the parameters used to deliver the vaccine equally to every individual? The difference and important element of this study from other studies is that it is not only prepared by conducting relevant research, but also provides institutions and organizations with the opportunity and example of implementation with indicator panels, and helps them make strategic decisions by seeing the positive and, if any, negative aspects of the application.

Contributors

Assoc. Prof. Dr. M. Hanefi CALP
Ankara Hacı Bayram Veli University
Faculty of Economics & Administrative Sciences
Department of Management Information Systems, Ankara, Türkiye
Email: hanefi.calp@hbv.edu.tr, **ORCID:** 0000-0001-7991-438X

Resul BÜTÜNER
Ministry of National Education,
Innovation and Educational Technologies General Directorate, Ankara, Türkiye
Email: rbutuner@gmail.com, **ORCID:** 0000-0002-9778-2349

Kartal Derin
SmartPro Technology, AI Department, İstanbul, Türkiye
Email:kartalderinmail@yahoo.com, **ORCID:** 0009-0009-7104-5955

Lecturer Dr. Akça Okan YÜKSEL
Middle East Technical University
Information System Coordination, Rectorship, Ankara, Türkiye
Email:akca@metu.edu.tr, **ORCID:** 0000-0002-5430-0821

Assist. Prof. Dr. Murat Kilinc
Karadeniz Technical University
Faculty of Economics & Administrative Sciences
Department of Management Information Systems, Trabzon, Türkiye
Email: muratkilinc@ktu.edu.tr, **ORCID:** 0000-0003-4092-5967

Dr. Hüseyin ULUKUZ
Ministry of National Education,
Innovation and Educational Technologies General Directorate, Ankara, Türkiye
Email: huseyin.ulukuz@hotmail.com , **ORCID:** 0000-0003-2272-6112

Research Assistant Dr. Murat ATASOY
Trabzon University
Faculty of Education
Computer and Instructional Technologies Education, Trabzon, Türkiye
Email:murat.atasoy@trabzon.edu.tr, **ORCID:** 0000-0001-6589-0161

Lecturer Dr. Adil YILDIZ
Trabzon University
Faculty of Education
Computer and Instructional Technologies Education, Trabzon, Türkiye
Email: adilyildiz@trabzon.edu.tr, **ORCID:** 0000-0002-7383-3885

Assist. Prof. Dr. Lokman ŞILBIR
Trabzon University
Çarşıbaşı Vocational School
Computer Technologies Department, Trabzon, Türkiye
Email: adilyildiz@trabzon.edu.tr, **ORCID:** 0000-0003-3655-2512

Assoc. Prof. Ekrem BAHÇEKAPILI
Karadeniz Technical University
Faculty of Economics & Administrative Sciences
Department of Management Information Systems, Trabzon, Türkiye
Email: ekrem.bahcekapili@ktu.edu.tr, **ORCID:** 0000-0002-7538-1712

Assist. Prof. Dr. Ramazan KOCAOĞLU
Ostim Technical University
Faculty of Engineering
Department of Computer Engineering, Ankara,Türkiye
Email: ramazan.kocaoglu@ostimteknik.edu.tr, **ORCID:** 0000-0002-6554-3335

Assist. Prof. Dr. Özgür TONKAL
Samsun University
Faculty of Engineering and Natural Sciences
Departman of Software Engineering, Samsun, Türkiye
Email: ozgur.tonkal@samsun.edu.tr, **ORCID:** 0000-0001-7219-9053

Assist. Prof. Dr. Erdal ÖZDOĞAN
Uludağ University
Faculty of Business Administration, Department of Management Information Systems, Inegol, Bursa, Türkiye
Email: erdalozdogan@uludag.edu.tr , **ORCID:** 0000-0002-3339-0493

Dr. Onur CERAN
Gazi University
IT Departmant, Ankara, Türkiye
Email: onur.ceran@gazi.edu.tr, **ORCID:** 0000-0003-2147-0506

Assoc. Prof. Dr. Mehmet KOKOÇ
Trabzon University
School of Applied Sciences
Department of Management Information Systems, Trabzon, Türkiye
Email: kokoc@trabzon.edu.tr, **ORCID:** 0000-0002-1347-8033

Data-Driven Analytical Techniques in Geographic Information Systems

Murat KILINÇ

Karadeniz Technical University

To Cite This Chapter

Kılınç, M. (2024). Data-Driven Analytical Techniques in Geographic Information Systems. In M. Hanefi Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 1-19). ISRES Publishing.

Introduction

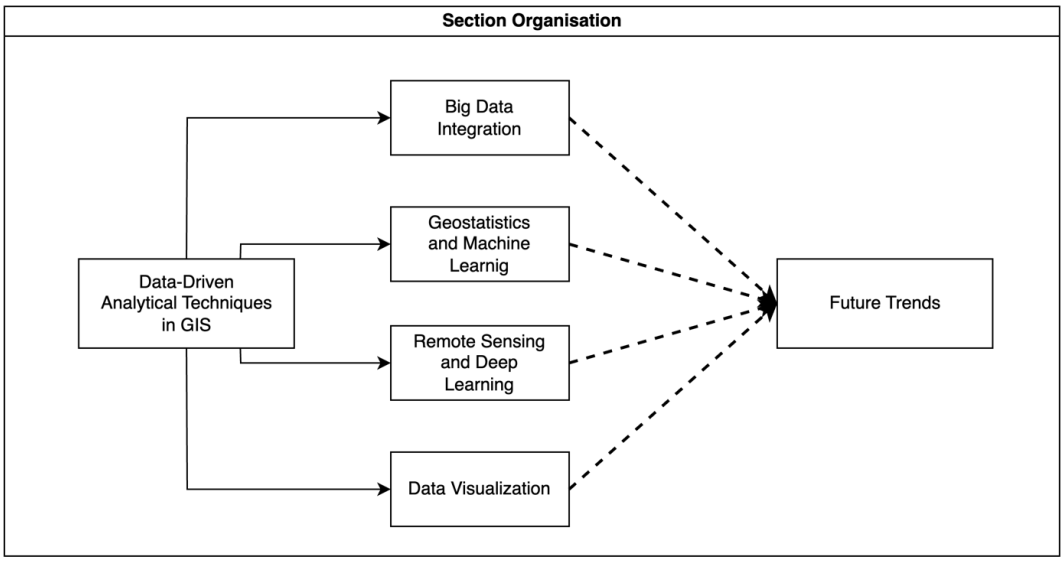
Geographic Information Systems (GIS) have become indispensable tools used in different disciplines to analyze and manage spatial data. GIS applications have an important place in many areas such as urban planning, environmental monitoring, transportation, and disaster management. The rapid increase in the amount of spatial data from satellites, sensors, and mobile devices has increased the need for advanced analytical techniques to make sense of these large datasets (Villarroya et al., 2022). In recent years, data-driven approaches have emerged as powerful methods that enhance the capabilities of GIS and have assumed an important role in more accurate predictions, in-depth analyses, and effective decision-making processes (Reichstein et al., 2019). Data-driven analytical techniques analyze spatial data using the power of data science, machine learning, artificial intelligence, and statistical methods. These techniques offer new perspectives that traditional GIS methods may not be able to capture by identifying patterns, trends, and relationships within large datasets. In this context, through the integration of data science with GIS, organizations can explore new dimensions in spatial data analysis, optimize resource allocation, and find solutions to complex spatial problems that could not be solved before (Cao et al., 2024).

Historically, GIS was mostly used for mapping and basic spatial analyses and was developed based on simple geographic techniques with manually collected data (Ahasan & Hossain, 2021). However, with the increase in high-resolution spatial data obtained through satellites, drones, and other remote sensing technologies, GIS has become a more dynamic and complex system. Today, GIS is not limited to the production of static maps but is used in a much wider range of applications such as real-time analysis, predictive modeling, and data visualization. This evolution has accelerated with the merger of GIS and data science. As spatial datasets have increased in both size and complexity, traditional GIS tools have struggled to manage and analyze these large information sources. Data-driven techniques such as machine learning, deep learning, and advanced statistical methods have filled this gap by providing powerful tools to process, analyze, and make sense of large-scale spatial data. These methods enable GIS professionals to gain valuable insights, model future scenarios, and make data-driven decisions (Li, Zhao, et al., 2022). Accordingly, big data plays a central role in modern GIS applications. Big data refers to data sets that are too large or complex to be managed by traditional data processing methods (Piovani & Bonovas, 2022). In the context of GIS, big data comes from various sources such as satellite imagery, sensor networks, mobile devices, social media, and Internet of Things (IoT) devices. These data streams provide extensive information about the physical world, such as land use patterns, weather, transport

flows, and environmental changes. Thus, data-driven analytical techniques enable more comprehensive and accurate analyses using big data. In particular, combining spatial data with other types of data, such as demographic, economic, or environmental information, can help scientists develop models that reflect the complexity of real-world systems. For example, in urban planning, big data can be used to analyze traffic flows, predict congestion hotspots, and optimize public transport routes. In environmental monitoring, big data can be used to monitor deforestation, assess the impact of climate change, and model the spread of pollutants. In this direction, big data also enables real-time analyses in GIS. This is especially critical for disaster management and emergency response applications (Sarker et al., 2020; Shah et al., 2019). By analyzing live data from sensors, drones, and other devices, authorities can monitor evolving situations such as floods or forest fires and make timely decisions to minimize damage. One of the most important developments in data-driven GIS is the integration of machine learning algorithms. Machine learning is a sub-branch of artificial intelligence that enables computers to learn from data and make predictions or decisions without being explicitly programmed (Yue et al., 2020). In GIS, machine learning algorithms are used to detect patterns, classify data, and make predictions based on spatial data. For example, machine learning models can be trained to classify land cover types from satellite imagery, detect urban growth patterns, or predict the probability of natural disasters such as floods or landslides. These models can process large amounts of spatial data and identify complex patterns that are difficult or impossible to detect manually by humans. Furthermore, machine learning can also be used for predictive modeling in GIS; this includes applications such as predicting future land use changes, forecasting population growth or modeling the impact of climate change on ecosystems.

Figure 1

Data-Driven Analytical Techniques in GIS Schema



The integration of data-driven techniques with GIS increases the accuracy of spatial analyses, accelerates decision-making processes, and plays an active role in solving wider spatial problems. With these technologies, cities can be planned more efficiently, natural resources can be better protected and the impact of environmental changes can be monitored more accurately. As a result, this evolution in GIS and data science is opening new horizons in the world of spatial analysis and helping to shape the cities, environment, and infrastructure of the future. Accordingly, this research focuses on the basic components and application areas of data-driven analytical techniques in the context of Geographic Information Systems (GIS). In this context, data-driven analytical techniques in the field of GIS are examined, and research on future trends is presented

by making in-depth analyses of big data integration, geo statistics, remote sensing, deep learning, and data visualization (Figure 1). The study also discusses the challenges faced in data-driven GIS analyses and potential future research areas.

Foundations of Data-Driven GIS Analytics

Geographic Information Systems (GIS) are comprehensive tools for analyzing, managing, and visualizing spatial data. GIS plays a critical role in many areas such as urban planning, natural resource management, transport systems, environmental monitoring, and disaster management. Traditional uses of GIS were limited to mapping and basic spatial analyses, but in recent years, with the development of data science and big data analysis, GIS has become more advanced and dynamic. Data-driven analytical techniques have significantly expanded the potential of GIS for spatial data analysis and provided powerful tools for making sense of large data sources. The fundamentals of data-driven GIS analysis involve the combination of disciplines such as data science, statistics, machine learning, and artificial intelligence to analyze spatial datasets. These analytical approaches enable deeper insights, more accurate predictions, and optimized decision-making processes on spatial data. Especially the use of big data sources allows us to understand why data-driven analytical techniques have become so important in GIS.

Fundamentals of GIS and Spatial Data

Geographic Information Systems is an integrated software platform used to collect, manage, analyze, and visualize spatial data (Eccles et al., 2019). Spatial data is a type of data that expresses the location of an object or event and is usually defined by geographic coordinates. These data are used for mapping and spatial analysis. The foundations of GIS are based on vector and raster data types. Vector data are expressed in geometric shapes such as points, lines, and polygons, while raster data are continuous data types consisting of pixels such as satellite images or digital elevation models. Traditional GIS tools were used to process this data and perform basic mapping operations. However, in today's big data era, the huge amount of spatial data from various sources such as satellite imagery, sensors, mobile devices, and social media requires more complex analysis techniques. This is where data science and data-driven analytical techniques come into play.

Data Science and GIS Integration

Data science is an interdisciplinary field that aims to obtain meaningful insights by analyzing large data sets and includes various methods such as statistics, data mining, machine learning, and artificial intelligence. The integration of data science with GIS allows spatial data to be analyzed in a more complex way and to obtain more efficient results. Data-driven GIS analyses focus on discovering patterns, trends, and spatial relationships by processing spatial data collected from large data sources. For example, in urban planning, data science methods can help make more efficient planning decisions by analyzing numerous factors such as traffic flows, population densities, and environmental changes. Similarly, in environmental monitoring applications, more accurate predictions can be made using big data sources to analyze the effects of climate change. Data science, combined with GIS, contributes to solving broader problems by increasing the accuracy of spatial analyses.

The Role of Big Data

Big data refers to data sets that are too large and complex to be managed by traditional methods. Today, GIS is an important tool for analyzing the huge amount of spatial data coming from big data sources. Big data sources are usually collected from sensor networks, satellite imagery, mobile devices, and social media platforms. These data

provide a large knowledge base that can be used for spatial analyses. Data-driven GIS analyses provide the integration of big data and GIS. This integration helps to make better decisions in various application areas, from urban planning to environmental monitoring. For example, satellite imagery, when combined with data from weather sensors, can lead to more accurate and timely decisions in disaster management. Similarly, traffic data from mobile devices can be used to optimize urban transport systems.

Machine Learning and GIS

Machine learning is a technique frequently used in data-driven GIS analyses. Machine learning enables computers to learn from specific datasets and make future predictions or decisions. The combination of GIS and machine learning is highly effective for detecting complex patterns in spatial data and building predictive models. Machine learning is used in various GIS applications such as the classification of satellite imagery, detection of land use changes, and prediction of natural disaster risks. By processing large amounts of spatial data, these algorithms can reveal complex relationships that humans cannot detect manually. In addition, machine learning models can also be used to model future scenarios and make predictions in various fields.

Geostatistics and GIS

Geostatistics is another important method used to perform statistical analyses on spatial data. Geostatistical methods in GIS are used to analyze the distribution of spatial variables in a given region and to understand the relationships between these variables. Geostatistical models, such as kriging, are one of the widely used techniques for predicting and modeling spatial data. Geostatistical methods play an important role, especially in environmental analyses and natural resource management. For example, geostatistics is used to estimate the productivity of agricultural areas or to model the distribution of water resources in a region. Such methods, when combined with GIS, allow for more accurate spatial analyses.

Data Visualisation and GIS

Data visualization is an important component of GIS analyses. Visualization of spatial data allows users to better understand the data and interpret the results of analysis more effectively. Maps, graphs, and 3D models are common tools used in the visualization of spatial data. Data visualization is especially important in decision-making processes. By visualizing the results of spatial analyses, decision-makers can better understand what kind of changes they need to make in a particular area. For example, urban planners can visualize data to map how busy certain roads are to optimize traffic flows. Similarly, disaster management experts can visualize risk zones to analyze the effects of natural disasters.

Big Data Integration and Spatial Data Analysis Techniques

Big data integration has revolutionized spatial data analysis, increasing the efficiency and effectiveness of modern Geographic Information Systems (GIS) (Al-Yadumi et al., 2021; Huang & Wang, 2020; Werner, 2019). GIS is used in many sectors as a powerful tool for the collection, analysis, and visualization of spatial data. GIS, which has a wide range of applications such as urban planning, natural resource management, transport, environmental monitoring, and disaster management, has become capable of solving more complex problems by being supported by big data sources. The increase in data from satellite images, sensor data, social media, and mobile devices has led to the necessity of integrating big data analysis techniques with GIS. In this integration process, various software that provides big data management and analysis (Table 1) are combined with GIS to enable more comprehensive spatial analyses. Platforms such as Apache Hadoop and Apache Spark are prominent in storing and processing large data sets. They are

particularly suitable for processing high-volume data sources such as satellite imagery and sensor data. For example, Apache Spark's distributed processing capabilities enable data sets to be analyzed quickly and monitor traffic density or environmental changes in cities. Without such data processing tools, performing spatial analyses on large datasets would be difficult and time-consuming.

Cloud-based platforms also play an important role in GIS and big data integration. Solutions such as Google BigQuery and Amazon Web Services (AWS) enable large data sets to be quickly stored and analyzed in the cloud environment (Al-Yadumi et al., 2021). For example, studies such as population density and traffic analysis in urban planning projects can be carried out based on data stored in such cloud-based systems. While these platforms accelerate data analysis with SQL-like queries, they also facilitate remote access to data by users. Solutions for the integration of GIS software with big data include platforms such as ESRI ArcGIS and Google Earth Engine. ESRI ArcGIS enables visualization and analysis of large volumes of geographic data by providing strong integration with big data through modules such as GeoAnalytics Server (Mai et al., 2019). Google Earth Engine, on the other hand, focuses on analyzing satellite data and provides a powerful platform for monitoring environmental changes and urban growth analyses. Such platforms not only accelerate data analysis but also enable visual analyses so that decision-makers can gain deeper insights into spatial data. Database extensions such as the open-source GIS platform QGIS and PostGIS have a wide range of uses in big data analyses. QGIS offers flexibility in spatial analyses by integrating with many data sources. PostGIS, on the other hand, allows spatial queries on large data sets, which enables high-performance data processing, especially in areas such as urban data management and environmental monitoring (Janisio-Pawłowska & Pawłowski, 2024). Open-source solutions are often preferred in small and medium-sized projects due to their cost advantage and flexibility. Tools such as GeoMesa and GeoSpark, which work with distributed databases, are ideal for spatial analysis on large datasets. GeoMesa is compatible with distributed databases such as Apache Accumulo, Cassandra, and HBase when performing spatial analysis on large databases. This can be used in areas such as real-time monitoring of sensor data. GeoSpark, on the other hand, runs on Apache Spark, enabling high-speed data processing in areas such as traffic density analysis and disaster area analysis. Such tools are especially advantageous in dynamic environments with continuous data flow (Dritsas et al., 2020).

Real-time analysis and visualization tools are also an important part of big data and GIS integration. Software such as Tableau and MapD (OmniSci) offer powerful solutions for spatial data visualization. For example, Tableau provides a user-friendly interface for urban planning, disaster management, and visualization of environmental changes through big data integration (Bivand, 2022). These tools facilitate the understanding of spatial data through maps, graphs, and interactive visuals. MapD, which offers GPU-accelerated data processing, stands out with its real-time data processing performance in environmental monitoring and traffic density analyses. Advanced analysis techniques such as machine learning and deep learning further advance GIS applications with big data integration. In big data analysis, machine learning techniques provide important contributions in areas such as predicting crime rates in cities or predicting traffic density by creating predictive models. Deep learning is used in projects such as disaster management and environmental change monitoring that require more complex data analyses (Sun & Scanlon, 2019). For example, deep learning models trained on satellite imagery can analyze the spatial distribution of forest fires and provide rapid intervention to the authorities. Such techniques allow for more accurate predictions when working on large data sets. Finally, database solutions such as IBM Db2 Big SQL enable spatial data analysis through SQL-based queries with big data sources.

Table 1
Big Data Software Integrated into Geographic Information Systems and Features

Software / Tool	Scope	Features	Use Cases
Apache Hadoop	Big data storage and processing	Data storage with HDFS (Hadoop Distributed File System), data processing with MapReduce	Satellite data processing, sensor data storage, urban data analysis
Apache Spark	Big data analytics and processing	Distributed data processing, stream analytics, GIS data analysis with PySpark and Scala support	Real-time analytics, environmental monitoring, disaster management
Google BigQuery	Big data analysis and management	SQL-like queries for big data analysis, high-speed data processing	Urban planning, traffic density analysis, environmental data analysis
ESRI ArcGIS	GIS platform	Geographic data integration, data visualization, big data modules (GeoAnalytics Server)	Satellite imagery, urban planning, environmental monitoring, disaster management
Google Earth Engine	Satellite and environmental data analytics	Processing large satellite data, ready-to-use datasets for environmental analysis	Deforestation, water resources monitoring, urban growth analysis
Amazon Web Services (AWS) S3	Cloud data storage and big data management	Storage and processing of large datasets, API support for GIS software integration	Satellite data storage, real-time urban data processing
Azure Synapse Analytics	Big data integration and analytics	Distributed data processing, machine learning models for GIS data analysis	Environmental change monitoring, urban infrastructure analysis, disaster management
QGIS	Open-source GIS platform	Integration with big data modules, data extraction from various big data sources	Urban planning, environmental change analysis, geographic distribution modeling
PostGIS	Database extension for geographic data	Spatial queries on large datasets, data storage, high-performance geographic data processing	Urban data management, environmental monitoring, geographic data storage
GeoMesa	Spatial analysis for big data	Spatial analysis in distributed databases (compatible with Apache Accumulo, Cassandra, and HBase)	Sensor data analysis, real-time monitoring, disaster management
GeoSpark (Apache Sedona)	Spatial analysis on large datasets	Spatial data processing on Apache Spark, parallel GIS data analysis	Traffic density analysis, urban planning, disaster area analysis
Hortonworks Data Platform	Big data integration	GIS data processing on the Hadoop ecosystem, real-time analytics	Satellite imagery, sensor data, urban infrastructure
Tableau	Data visualization	Big data integration with spatial data visualization, map-based data analysis	Urban planning, disaster management, visualization of environmental changes
MapD (OmniSci)	GPU-accelerated big data analytics and GIS	Real-time spatial data processing and visualization, high-speed data querying	Environmental monitoring, disaster management, traffic density analysis
IBM Db2 Big SQL	Structured and unstructured big data analysis	SQL-based spatial data analysis on Hadoop and other data sources	Urban infrastructure analysis, environmental data integration, GIS analysis on big data

IBM Db2 Big SQL, which can be integrated with Hadoop and other big data sources, is used in projects such as urban infrastructure analysis and environmental data integration. Such solutions add value to GIS applications from a big data perspective by providing flexibility in analyzing both structured and unstructured data.

In conclusion, big data integration and spatial data analysis techniques are important developments that expand the potential of GIS and offer solutions to the complex spatial problems of the modern world. The increase in big data sources has made it necessary to go beyond traditional GIS methods and get support from fields such as data science, machine learning, and deep learning. Integration of these techniques with GIS enables more accurate analyses, faster decision-making processes, and solving wider spatial problems. In the future, deeper integration of big data and GIS will enable more complex analyses and more effective decisions in many areas from urban planning to environmental monitoring. Big data-driven GIS will help shape the cities, environment, and infrastructure of the future, contributing to the sustainable development of modern societies.

Geostatistics and Machine Learning in GIS

Geographical Information Systems (GIS) is a powerful tool that enables analyzing spatial data in many fields and drawing meaningful conclusions from these data. In order to use the power of GIS more effectively, different methods and techniques are utilized. Among these methods, geostatistics and machine learning techniques, which have a wide range of applications in data science, stand out. While geostatistics tries to understand spatial patterns and variability by analyzing the relationships between spatial data, machine learning provides the ability to learn data patterns and predict future events or situations. The harmonious use of both methods in GIS makes it possible to reach faster and more accurate solutions to complex spatial problems.

Geostatistical methods often examine the spatial distributions of data, allowing us to make predictions in missing or unsampled regions. These methods utilize statistical models to explore the trends and spatial dependencies of a given area. Geostatistical methods, which are frequently used in the analysis of environmental variables and management of natural resources, enable detailed analyses in many areas from environmental pollution to water quality (Hasan et al., 2021). The table below summarises the main geostatistical methods used in the context of GIS and the areas in which these methods are prominent (Table 2). When the table is examined in detail, Kriging, one of the geostatistics methods, can generate predictions for non-sampled locations by analyzing the spatial dependence between data points. This method is especially effective in applications such as determining the distribution of air pollution throughout the city and creating water quality maps (Miao & Wang, 2024). Thiessen Polygons creates regions specific to each data point and enables analyses based on the distance between these points. It is frequently used in precipitation measurements and population density analyses. IDW (Inverse Distance Weighting) determines the spatial distribution of environmental variables by giving more weight to data points closer to the location to be predicted; it is a preferred method for creating temperature maps and analyzing terrain features. Semivariogram analyses spatial dependence over distance, while Spline Interpolation is used to create a continuous surface from data points and is useful in analyses such as elevation maps (Li, Baorong, et al., 2022).

Spatial Autocorrelation analyses the distribution of similar values using Moran's I, thus contributing to the identification of spatial patterns such as water pollution or vegetation cover. Finally, Trend Surface Analysis allows us to understand large-scale spatial trends and is used to analyze issues such as land slope or temperature variations (Love et al., 2022).

Table 2
Geostatistical Methods for Spatial Data Analysis in GIS

Method	Description	Application Areas
Kriging	An interpolation technique for spatially estimating data points in geographic datasets	Air pollution distribution, water quality analysis, soil properties
Thiessen Polygons	Mapping and defining unique regions for each data point	Rainfall measurement data, population density analysis, service area determination
IDW (Inverse Distance Weighting)	Distributes values from points to surrounding areas, giving weights based on distance	Temperature maps, land properties, pollutant distribution
Semivariogram	Examines spatial variability and determines spatial relationships between data points	Soil moisture analysis, spatial analysis of climate variables
Spline Interpolation	Used to create a continuous surface from data points	Surface modeling, elevation maps, detailed terrain analysis
Spatial Autocorrelation (Moran's I)	Analyzes how data is distributed spatially and identifies similarities	Water pollution analysis, vegetation distribution, urban density analysis
Trend Surface Analysis	Used to understand trends within a specific area	Terrain slope and elevation analysis, temperature and rainfall trends

Table 3
Machine Learning Techniques for GIS Applications

Machine Learning Method	Description	Application Areas
Support Vector Machines (SVM)	Used for classification and regression analysis; finds the optimal hyperplane to separate data	Land classification, natural disaster risk analysis
Decision Trees	Classifies data by branching based on specific decision rules	Environmental risk analysis, water resources assessment
Random Forests	A model formed by combining multiple decision trees, commonly used for classification and regression	Land cover classification, deforestation analysis, urban development analysis
K-Nearest Neighbors (KNN)	Classifies data based on similarity between neighboring points; non-parametric method	Land type classification, population density analysis
Artificial Neural Networks (ANN)	Performs prediction and classification on complex datasets; consists of layers and learns from data	Air pollution prediction, environmental change monitoring, disaster risk assessment
Deep Learning	Advanced analysis and prediction on large datasets, especially in image analysis	Satellite image analysis, forest fire detection, traffic density prediction
Semi-Supervised Learning	Learning process with limited labeled data, useful for data constraints	Spatial data classification, urban data analysis
Clustering (K-Means)	Groups data into meaningful clusters by assigning each data point to the nearest cluster	Population density, natural resource distribution, land use analysis

Machine learning methods offer great advantages in terms of speed and accuracy in GIS analyses. These methods, which automate processes such as classification, prediction, and pattern recognition, especially on large datasets, provide valuable information in environments with complex and dynamic data. The table above summarises the main machine learning techniques commonly used for GIS and their application areas (Table 3).

Support Vector Machines (SVM) is a powerful method in classification and regression analyses and can find the most appropriate hyperplane to classify data into classes. This method provides effective results in land classification and natural disaster risk analyses (Yousefi et al., 2020). Decision Trees branch and classify data by applying certain decision rules at each node; they are widely used in environmental risk analyses and water resource assessment (Rodríguez et al., 2021; Sánchez-Ortiz et al., 2021). Random Forests provides a model that improves accuracy by combining multiple decision trees and is successful in land cover classification, deforestation analysis, and urban development analysis. K-nearest neighbor (KNN) classifies according to the similarity of neighboring points; it is especially preferred in land type classification and population density analysis (Ge et al., 2020). Artificial Neural Networks (ANN) perform prediction and classification on complex datasets and are widely used in areas such as air pollution prediction, environmental change monitoring, and disaster risk assessment. Deep Learning, on the other hand, has the capacity to perform advanced analyses on large datasets and plays an important role in satellite image analysis, forest fire detection, and traffic density prediction (Adegun et al., 2023). While Assisted Learning is used in areas such as data classification and urban data analysis by working with under-labeled data, Clustering (K-Means) groups data into meaningful clusters and is a preferred technique for population density, natural resource distribution, and land use analysis.

GIS offers a wide range of analyses and predictions by using both geostatistics and machine learning techniques in spatial data analysis. Geostatistics provides an important tool in addressing environmental and social issues by modeling spatial relationships between data. Machine learning, on the other hand, brings flexibility to GIS by extracting meaning from large data sets and automatic learning processes. When these two methods come together, versatile analysis opportunities are provided for users to make more accurate and reliable decisions. The impact of GIS is expanding in many areas from urban planning to natural resource management and disaster monitoring. Faster processing of data and detailed analyses create a strong infrastructure for early detection and management of environmental problems. Thus, complex spatial data gain meaning and strategic decision-making processes are based on more solid ground. These techniques, which increase the power of GIS, also contribute to the creation of more sustainable and smart cities in the future.

Remote Sensing and Deep Learning for Spatial Data

Remote sensing and deep learning are two powerful analysis tools that are becoming increasingly critical in modern Geographic Information Systems (GIS). With the development of technology, huge datasets can be collected to help us understand and observe our environment. Remote sensing provides high-resolution data from large areas with techniques such as satellite imagery, LiDAR, and thermal imaging, and is used in a wide range of fields from tracking environmental events to urban planning. Deep learning methods come into play in the analysis of these data; they can quickly and accurately perform operations such as extracting meaningful information from large data sets, pattern recognition, classification, and prediction. While remote sensing techniques increase the effectiveness of GIS projects by continuously collecting data from the environment, deep learning algorithms enable more advanced analyses by making use of these data (Zhang et al., 2022). The combination of the two approaches provides great benefits in a wide range of applications such as environmental monitoring, planning of urban areas, and natural disaster management.

Remote sensing techniques provide a wide data source in the process of collecting environmental and spatial data. The main advantage of these techniques is the ability to analyze different features by collecting information from large geographical areas. Satellite imagery allows continuous observation of large areas, while aerial photography allows more detailed data to be collected. LiDAR technology is used to create high-resolution 3D maps of the land surface using laser beams. Multispectral and hyperspectral imaging techniques allow the identification of features such as vegetation, water resources, and mineral distribution in a given area (Buckley et al., 2013). Thermal imaging is used in analyses such as the urban heat island effect by monitoring temperature changes, while radar remote sensing offers the advantage of obtaining images regardless of weather conditions. The following table summarises the remote sensing techniques commonly used in GIS applications and their application areas (Table 4).

Table 4
Remote Sensing Techniques in GIS

Technique	Description	GIS Applications
Satellite Imagery	High-resolution images captured from satellites for various spatial analyses	Land use classification, urban growth analysis, environmental monitoring
Aerial Photography	Images taken from aircraft to capture detailed spatial data	Topographic mapping, disaster assessment, infrastructure planning
LiDAR (Light Detection and Ranging)	Uses laser pulses to measure distances and create high-resolution 3D maps of the Earth's surface	Terrain modeling, forestry analysis, flood risk assessment
Multispectral Imaging	Captures images across multiple wavelengths to identify material characteristics	Vegetation analysis, water body detection, mineral exploration
Thermal Imaging	Uses infrared sensors to detect temperature variations in the environment	Urban heat island analysis, forest fire detection, industrial monitoring
Radar Remote Sensing	Uses radar waves to capture images, allowing for analysis regardless of weather conditions	Land deformation monitoring, soil moisture analysis, oceanographic studies
Hyperspectral Imaging	Captures a broad spectrum of light to distinguish materials based on their spectral signatures	Agricultural assessment, mineral identification, environmental monitoring

Satellite imagery makes it possible to monitor large areas quickly and with high resolution. Satellite images, which are preferred for land use classification and urban growth analyses, are also important for environmental monitoring studies. Aerial photography allows obtaining more detailed spatial data and is used in studies that require detailed analysis such as disaster assessment and infrastructure planning. LiDAR maps the land surface in high resolution with laser pulses. This technique is very effective in forest analyses, terrain modeling, and flood risk assessment. Multispectral imaging enables versatile analyses such as vegetation analysis, detection of water resources, and mineral exploration. Thermal imaging analyses environmental problems such as the heat island effect in urban areas and the detection of forest fires by measuring temperature changes. Radar remote sensing collects data regardless of weather conditions and finds applications in different fields such as land deformation monitoring and ocean research. Hyperspectral imaging offers the advantage of separating materials according to their spectral signatures in studies such as agricultural evaluation and mineral detection (Shukla & Kot, 2016).

Deep learning techniques play an important role in the process of analyzing and learning large data sets. It provides fast and accurate results by automating complex processes such as pattern recognition, classification, and prediction, especially in GIS data. Deep learning performs tasks such as detecting buildings, creating road maps, and classifying vegetation by performing detailed analyses on satellite and aerial images. In the table below, the main deep learning techniques used in GIS applications and the usage areas of these techniques are given (Table 5).

Table 5
Deep Learning Techniques in GIS

Technique	Description	GIS Applications
Convolutional Neural Networks (CNN)	Deep learning techniques specialized in image analysis and feature extraction	Land cover classification, urban structure detection, deforestation monitoring
Recurrent Neural Networks (RNN)	Processes sequential data, useful for time series spatial data analysis	Weather prediction, traffic flow analysis, temporal change detection
Generative Adversarial Networks (GAN)	Used for generating synthetic data and enhancing image quality	Satellite image enhancement, filling missing data, land use simulation
Autoencoders	Unsupervised learning method used for dimensionality reduction and feature extraction	Noise reduction in remote sensing data, anomaly detection in spatial datasets
Transfer Learning	Leveraging pre-trained models for similar tasks to reduce training time and improve accuracy	Rapid deployment for land cover classification, disaster assessment
Semantic Segmentation	Pixel-level image classification for detailed object detection	Building and road detection, vegetation mapping, land use analysis
Object Detection	Identifies specific objects within an image, used for counting and tracking	Tree counting, vehicle detection, wildlife monitoring

Convolutional Neural Networks (CNN) perform particularly well in image analysis and are used in areas such as land cover classification and urban structure detection. Recurrent Neural Networks (RNN) are a suitable technique for analyzing time series data and are preferred in weather forecasting or traffic flow analyses. Generative Adversarial Networks (GAN) are used to improve image quality and generate synthetic data. It is an ideal technique for processes such as satellite image enhancement and missing data filling (Zaytar & El Amrani, 2021). Autoencoders are used as an unsupervised learning method to reduce the dimensionality of data and feature extraction; it is useful for noise reduction in remote sensing data and anomaly detection in spatial data sets. Transfer learning reduces training time and improves accuracy by using pre-trained models on similar tasks, enabling rapid land cover classification and disaster assessment (Alem & Kumar, 2022). Semantic segmentation provides detailed object detection by classifying each pixel in the image and is used in detailed analyses such as building and road detection. Object detection, on the other hand, identifies specific objects in the image and is useful in counting and tracking operations; it is especially effective in tasks such as tree counting, vehicle detection, and wildlife monitoring.

In general terms, the combination of remote sensing and deep learning techniques provides in-depth analyses and fast solutions in GIS projects. These methods support

decision-making in critical areas such as environmental monitoring, urban planning, and disaster management by making sense of big data sources. With the developing technology, the efficiency and accuracy provided by these techniques contribute to the creation of more sustainable and smart cities. While remote sensing enables the collection of spatial data covering a wide area, deep learning enables making predictions for the future by making sense of these data. In this direction, the combination of these two methods in GIS applications has the potential to enable better management of modern cities and environmental systems.

Spatial Data Visualization and Interpretation

Spatial data visualization and interpretation is an important component of data analysis in Geographic Information Systems (GIS). Visualization of spatial data supports users' decision-making processes by facilitating the understanding of complex geographic information. Visualization enables faster and more effective interpretation of data by representing multidimensional data with graphs, maps, and various visual tools. This contributes to the understanding of not only location-based information but also patterns, relationships, and variability in the analysis of spatial data. In this context, data visualization in GIS has a wide range of applications from urban planning to natural resource management, from environmental monitoring to crisis and disaster management (Nasr-Azadani et al., 2023).

Data visualization tools and libraries enable GIS analysts to present spatial data in dynamic and interactive formats. For example, ArcGIS and QGIS are powerful GIS software widely used for spatial analysis and visualization. These platforms help urban planners analyze infrastructure development projects, environmental scientists monitor ecosystem changes, and public institutions to plan post-disaster recovery processes (Pavelka & Landa, 2024). The detailed analysis and mapping tools provided by ArcGIS allow users to view and analyze data in layers. For example, ArcGIS is often preferred for identifying natural disaster risk zones or monitoring urban expansion. QGIS, on the other hand, is a popular choice especially for research and academic studies due to its open-source nature and provides flexibility in spatial data analysis and mapping processes.

Table 6
Tools and Libraries for Spatial Data Visualization and Interpretation

Tool/ Library	Description	Areas of Use
ArcGIS	An advanced GIS platform for spatial analysis and data visualization	Mapping, geographic analysis, data visualization
QGIS	An open source GIS software; offers spatial data management and analysis capabilities	Map creation, data integration, analysis
Tableau	Data visualization and analysis platform; supports visualization of spatial data	Spatial data analysis, map-based data visualization
D3.js	JavaScript library for creating web-based data visualizations	Dynamic map visualizations, interactive data visualization
Leaflet	A lightweight JavaScript library for creating web-based interactive maps	Web mapping, interactive map creation
Mapbox	A web-based mapping platform with extensive map editing and visualization capabilities	Map creation, geographic data visualization, location-based analysis

Google Earth Engine	A cloud-based platform for analysis and visualization on large spatial datasets	Satellite imagery analysis, environmental change monitoring, spatial analysis
Matplotlib (Python)	PA Python-based data visualization library; can visualize GIS data in 2D	Graph drawing, data analysis, basic map visualization
GeoPandas (Python)	Python-based library for analyzing and visualizing geographic data	Geographic data processing, map analysis, spatial data visualization
Kepler.gl	A powerful tool for interactively visualizing large data sets	Geographic data visualization, urban planning, logistics analysis

Tableau, another important tool for data visualization, allows visualizing various data types, including spatial data. Tableau helps users explore spatial relationships within data by creating interactive maps. For example, spatial analyses of different data, such as population density, traffic flow, or health data, can be quickly and clearly visualized using Tableau (Taylor et al., 2021). These visualizations enable decision-makers to promptly understand the data and make strategic decisions with the visualized information. JavaScript libraries such as D3.js and Leaflet, which are used for web-based data visualization, enable GIS data to be shared with wider audiences. While D3.js is used to create dynamic and customizable visualizations, Leaflet is ideal for developing interactive maps thanks to its lightweight structure and user-friendly interface. For example, a Leaflet can be used to visualize air pollution rates in a city, allowing users to see the pollution density on a map. Such web-based visualizations enable a wide range of users to access spatial analyses and present data more effectively (Table 6).

Figure 2
Mapbox Usage Example

Platforms such as Mapbox offer flexibility in spatial data visualization, enabling the creation of high-resolution maps and the addition of customized data layers. For example, a logistics company can use Mapbox to visualize delivery routes (Figure 2) and reduce costs through route optimization.

Cloud-based solutions such as Google Earth Engine are used to analyze large spatial datasets. Google Earth Engine, which allows analyses of satellite images and environmental data, is particularly effective in monitoring and analyzing environmental changes. For example, a researcher who wants to analyse the deforestation process in Amazon forests can observe long-term changes by using the data sets provided by Google Earth Engine. In the Python ecosystem, libraries such as Matplotlib and GeoPandas are frequently preferred tools by data scientists and GIS experts for data analysis (Fleischmann et al., 2022). Matplotlib offers the user flexibility in creating 2D graphics and maps and is an ideal tool for simple maps. GeoPandas, on the other hand, is a powerful library for processing and visualizing geographic data and has a wide range of applications, especially in spatial data analysis. For example, an analyst who wants to examine the distribution of green space in a city can easily map different types of green space and their spatial relationships with GeoPandas. In addition, Kepler.gl, which is used for big data visualization, stands out as a powerful tool for presenting geographic data in an interactive and dynamic way (Zuo et al., 2022). Kepler.gl helps urban planners to analyze traffic density or logistics distribution networks. For example, an administrator who wants to monitor the traffic flow in a city can use Kepler.gl to visualize heavy traffic zones and make strategic decisions for necessary adjustments.

These tools and libraries enable spatial data to be presented in a more understandable and effective way. Visualization in GIS projects contributes directly to decision-making processes by making analysis results accessible and interpretable by a wider range of users. Visualization techniques allow the simplification of complex data structures and facilitate a better understanding of spatial patterns. The visual richness provided by these tools contributes to the more effective use of spatial analyses in a wide range of fields from academic research to industrial applications.

Challenges and Future Trends in Data-Driven GIS

Data-driven Geographic Information Systems (GIS) have become an important tool for collecting, analyzing, and interpreting spatial data, empowered by technologies such as big data analytics, artificial intelligence, and IoT. However, the development and implementation of data-driven GIS solutions involve many challenges. Various factors ranging from data quality to computational costs can affect the use and accuracy of these systems. At the same time, the predicted trends for the future development of GIS promise the emergence of smarter and more powerful systems. In this context, it is important to conduct a review of the challenges and future trends facing data-driven GIS solutions.

Challenges in Data-Driven GIS

Data-driven GIS requires a robust infrastructure to process and analyze large amounts of spatial data. However, the increase in the amount of this data and the increasing complexity of the data pose some significant challenges to GIS solutions. These challenges include data quality and accuracy, data integration and harmonization, data security and privacy, high computational costs, and data processing infrastructure.

Data Quality and Accuracy: The accuracy of data in GIS solutions directly affects the reliability of the results obtained. Data from different sources may sometimes be inconsistent or outdated. For example, incompatibilities between different data sources, such as satellite imagery or sensor data, can cause deterioration in data quality. Furthermore, some data may be incomplete or inaccurate, which can compromise the accuracy of GIS analyses. Therefore, it is important to develop automated error detection and data cleaning processes to improve data quality.

Data Integration and Harmonisation: In data-driven GIS solutions, integrating and

harmonizing data from different sources poses a major challenge. Various data types such as satellite images, GPS data, social media data, and sensor data from IoT devices need to be analyzed together. However, collecting these data in different formats and at different time intervals complicates the integration process. This situation requires harmonization of data in order to make accurate analyses in GIS solutions.

Data Security and Privacy: GIS projects are responsible for protecting data security and privacy, especially in studies involving large amounts of user data. Data containing location information of users, such as GPS data, raises privacy issues. Especially by combining data collected from security cameras in cities and social media platforms, there may be the potential to interfere with the private lives of individuals. Therefore, advanced encryption methods and data anonymization techniques should be used to ensure data security and privacy.

High Computational Costs: One of the most important challenges of data-driven GIS solutions is that big data processing requires high computational power and cost. Especially processing and analyzing large volumes of spatial data in real-time requires powerful hardware and advanced software. This increases the cost of GIS projects and makes it difficult for smaller-scale projects to access these technologies. While cloud computing solutions have the potential to reduce some of these high costs, cloud computing costs may also increase in the long term.

Data Processing Infrastructure and Performance: The data processing infrastructure required for GIS projects varies depending on the size of the data and the depth of the analysis. Especially in spatial analyses in big cities or time-critical projects such as monitoring natural disasters, fast and efficient data processing infrastructure is required. If a high-performance computing infrastructure is not established or is inadequate, data processing time may be prolonged and the accuracy of analyses may be adversely affected. Therefore, it is very important to use powerful and scalable data processing infrastructures in GIS projects.

Future Trends in Data-Driven GIS

The future of data-driven GIS has great potential with developing technologies. Innovations in areas such as artificial intelligence, machine learning, IoT, and cloud computing are making GIS solutions smarter, faster, and more effective. Future trends in data-driven GIS include smart city solutions, augmented reality (AR)-based mapping, real-time data analytics, more powerful AI-powered models, and cloud-based GIS solutions.

Smart City Solutions: Data-driven GIS solutions are critical for smart cities. Many aspects of city life such as traffic management, energy distribution, waste management, and security can be monitored and optimized with GIS. GIS solutions for smart cities contribute to making cities more sustainable and efficient. For example, thanks to real-time data from sensors, traffic density can be analyzed and transport routes can be managed more efficiently.

Augmented Reality (AR) and GIS: Augmented reality technology, when combined with GIS, can offer more interactive mapping and data presentation. Augmented reality-supported GIS solutions enable users to visualize environmental data more effectively. For example, it becomes possible to observe the status of infrastructure projects in a city or changes in natural disaster areas in real-time. This technology facilitates the use of GIS data in the field and provides users with a richer experience.

Real-Time Data Analytics: In the future, data-oriented GIS solutions are expected to become more effective with real-time data analytics. Thanks to IoT devices, satellite

systems, and sensor networks, real-time data can be collected continuously. These data can be quickly analyzed in GIS systems and provide instant information to decision-makers. Especially in areas that require instant intervention such as natural disaster management, urban traffic control, and environmental monitoring, real-time data analytics provide a great advantage.

Artificial Intelligence Supported GIS Models: Artificial intelligence and machine learning allow for more in-depth analysis of GIS data. In the future, AI-supported GIS solutions will be more widely used in areas such as predicting environmental risks, monitoring urban development, and modeling the effects of natural disasters. These models support decision-making processes and increase the effectiveness of GIS by producing faster and more accurate results. GIS systems integrated with artificial intelligence make significant contributions to data-driven decision-making processes by performing better in big data analyses.

Cloud-Based GIS Solutions: Cloud computing plays an important role in the future of data-driven GIS solutions. In projects with intensive big data processing, cloud-based GIS solutions facilitate data storage, processing, and analysis processes. Thanks to cloud technology, GIS projects can be supported with a cost-effective and flexible infrastructure. This enables small-scale businesses and local governments to access data-driven GIS solutions at lower costs.

The future of data-driven GIS enables the emergence of more advanced, intelligent, and user-friendly systems with technological innovations. GIS solutions powered by new technologies such as artificial intelligence and machine learning play an important role in solving environmental and social problems. New approaches such as augmented reality, real-time data analytics, and cloud-based solutions expand the usage areas of GIS and make it applicable in more sectors. These developments contribute to the adoption of data-driven GIS solutions by more institutions and organizations.

References

Adegun, A. A., Viriri, S., & Tapamo, J.-R. (2023). Review of deep learning methods for remote sensing satellite images classification: Experimental survey and comparative analysis. *Journal of Big Data*, *10*(1), 93. https://doi.org/10.1186/s40537-023-00772-x

Ahasan, R., & Hossain, M. M. (2021). Leveraging GIS and spatial analysis for informed decision-making in COVID-19 pandemic. *Health Policy and Technology*, *10*(1), 7–9. https://doi.org/10.1016/j.hlpt.2020.11.009

Alem, A., & Kumar, S. (2022). Transfer Learning Models for Land Cover and Land Use Classification in Remote Sensing Image. *Applied Artificial Intelligence*, *36*(1), 2014192. https://doi.org/10.1080/08839514.2021.2014192

Al-Yadumi, S., Xion, T. E., Wei, S. G. W., & Boursier, P. (2021). Review on Integrating Geospatial Big Datasets and Open Research Issues. *IEEE Access*, *9*, 10604–10620. https://doi.org/10.1109/ACCESS.2021.3051084

Bivand, R. (2022). R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data. *Geographical Analysis*, *54*(3), 488–518. https://doi.org/10.1111/gean.12319

Buckley, S. J., Kurz, T. H., Howell, J. A., & Schneider, D. (2013). Terrestrial lidar and hyperspectral data fusion products for geological outcrop analysis. *Computers & Geosciences*, *54*, 249–258. https://doi.org/10.1016/j.cageo.2013.01.018

Cao, K., Zhou, C., Church, R., Li, X., & Li, W. (2024). Revisiting spatial optimization in the era of geospatial big data and GeoAI. *International Journal of Applied*

Earth Observation and Geoinformation, *129*, 103832. https://doi.org/10.1016/j.jag.2024.103832

Dritsas, E., Kanavos, A., Trigka, M., Vonitsanos, G., Sioutas, S., & Tsakalidis, A. (2020). Trajectory Clustering and k-NN for Robust Privacy Preserving k-NN Query Processing in GeoSpark. *Algorithms*, *13*(8), 182. https://doi.org/10.3390/a13080182

Eccles, K. M., Pauli, B. D., & Chan, H. M. (2019). The Use of Geographic Information Systems for Spatial Ecological Risk Assessments: An Example from the Athabasca Oil Sands Area in Canada. *Environmental Toxicology and Chemistry*, *38*(12), 2797–2810. https://doi.org/10.1002/etc.4577

Fleischmann, M., Feliciotti, A., & Kerr, W. (2022). Evolution of Urban Patterns: Urban Morphology as an Open Reproducible Data Science. *Geographical Analysis*, *54*(3), 536–558. https://doi.org/10.1111/gean.12302

Ge, G., Shi, Z., Zhu, Y., Yang, X., & Hao, Y. (2020). Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: Performance assessment of four machine learning algorithms. *Global Ecology and Conservation*, *22*, e00971. https://doi.org/10.1016/j.gecco.2020.e00971

Hasan, K., Paul, S., Chy, T. J., & Antipova, A. (2021). Analysis of groundwater table variability and trend using ordinary kriging: The case study of Sylhet, Bangladesh. *Applied Water Science*, *11*(7), 120. https://doi.org/10.1007/s13201-021-01454-w

Huang, B., & Wang, J. (2020). Big spatial data for urban and environmental sustainability. *Geo-Spatial Information Science*, *23*(2), 125–140. https://doi.org/10.1080/10095020.2020.1754138

Janisio-Pawłowska, D., & Pawłowski, W. (2024). Implementation of BIM Data in CityGML—Research and Perspectives for Creating a QGIS Plugin for Spatial Analysis: Experience from Poland. *Sustainability*, *16*(2), 642. https://doi.org/10.3390/su16020642

Li, Y., Baorong, Z., Xiaohong, X., & Zijun, L. (2022). Application of a semivariogram based on a deep neural network to Ordinary Kriging interpolation of elevation data. *PLOS ONE*, *17*(4), e0266942. https://doi.org/10.1371/journal.pone.0266942

Li, Y., Zhao, Q., & Zhong, C. (2022). GIS and urban data science. *Annals of GIS*, *28*(2), 89–92. https://doi.org/10.1080/19475683.2022.2070969

Love, C. A., Skahill, B. E., Russell, B. T., Baggett, J. S., & AghaKouchak, A. (2022). An Effective Trend Surface Fitting Framework for Spatial Analysis of Extreme Events. *Geophysical Research Letters*, *49*(11), e2022GL098132. https://doi.org/10.1029/2022GL098132

Mai, G., Janowicz, K., Yan, B., & Scheider, S. (2019). Deeply integrating Linked Data with Geographic Information Systems. *Transactions in GIS*, *23*(3), 579–600. https://doi.org/10.1111/tgis.12538

Miao, C., & Wang, Y. (2024). Interpolation of non-stationary geo-data using Kriging with sparse representation of covariance function. *Computers and Geotechnics*, *169*, 106183. https://doi.org/10.1016/j.compgeo.2024.106183

Nasr-Azadani, E., Wardrop, D. H., & Brooks, R. P. (2023). Pathways for the utilization of visualization techniques in designing participatory natural resource policy and management. *Journal of Environmental Management*, *333*, 117407. https://doi.org/10.1016/j.jenvman.2023.117407

Pavelka, K., & Landa, M. (2024). Using Virtual and Augmented Reality with GIS Data. *ISPRS International Journal of Geo-Information*, *13*(7), 241. https://doi.org/10.3390/ijgi13070241

Piovani, D., & Bonovas, S. (2022). Real World—Big Data Analytics in Healthcare. *International Journal of Environmental Research and Public Health*, *19*(18), 11677. https://doi.org/10.3390/ijerph191811677

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rodríguez, D. J., Paltán, H. A., García, L. E., Ray, P., & St. George Freeman, S. (2021). Water-related infrastructure investments in a changing environment: A perspective from the World Bank. *Water Policy*, *23*(S1), 31–53. https://doi.org/10.2166/wp.2021.265

Sánchez-Ortiz, A., Mateo-Sanz, J. M., Nadal, M., & Lampreave, M. (2021). Water stress assessment on grapevines by using classification and regression trees. *Plant Direct*, *5*(4), e00319. https://doi.org/10.1002/pld3.319

Sarker, M. N. I., Peng, Y., Yiran, C., & Shouse, R. C. (2020). Disaster resilience through big data: Way to environmental sustainability. *International Journal of Disaster Risk Reduction*, *51*, 101769. https://doi.org/10.1016/j.ijdrr.2020.101769

Shah, S. A., Seker, D. Z., Hameed, S., & Draheim, D. (2019). The Rising Role of Big Data Analytics and IoT in Disaster Management: Recent Advances, Taxonomy and Prospects. *IEEE Access*, *7*, 54595–54614. https://doi.org/10.1109/ACCESS.2019.2913340

Shukla, A., & Kot, R. (2016). An Overview of Hyperspectral Remote Sensing and its applications in various Disciplines. *IRA-International Journal of Applied Sciences (ISSN 2455-4499)*, *5*(2), 85. https://doi.org/10.21013/jas.v5.n2.p4

Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters*, *14*(7), 073001. https://doi.org/10.1088/1748-9326/ab1b7d

Taylor, J. R., Hanumappa, M., Miller, L., Shane, B., & Richardson, M. L. (2021). Facilitating Multifunctional Green Infrastructure Planning in Washington, DC through a Tableau Interface. *Sustainability*, *13*(15), 8390. https://doi.org/10.3390/su13158390

Villarroya, S., Viqueira, J. R. R., Cotos, J. M., & Taboada, J. A. (2022). Enabling Efficient Distributed Spatial Join on Large Scale Vector-Raster Data Lakes. *IEEE Access*, *10*, 29406–29418. https://doi.org/10.1109/ACCESS.2022.3157405

Werner, M. (2019). Parallel Processing Strategies for Big Geospatial Data. *Frontiers in Big Data*, *2*, 44. https://doi.org/10.3389/fdata.2019.00044

Yousefi, S., Pourghasemi, H. R., Emami, S. N., Pouyan, S., Eskandari, S., & Tiefenbacher, J. P. (2020). A machine learning framework for multi-hazards modeling and mapping in a mountainous area. *Scientific Reports*, *10*(1), 12144. https://doi.org/10.1038/s41598-020-69233-2

Yue, P., Gao, F., Shangguan, B., & Yan, Z. (2020). A machine learning approach for predicting computational intensity and domain decomposition in parallel geoprocessing. *International Journal of Geographical Information Science*, *34*(11), 2243–2274. https://doi.org/10.1080/13658816.2020.1730850

Zaytar, M. A., & El Amrani, C. (2021). Satellite image inpainting with deep generative adversarial neural networks. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *10*(1), 121. https://doi.org/10.11591/ijai.v10.i1.pp121-130

Zhang, X., Zhou, Y., & Luo, J. (2022). Deep learning for processing and analysis of

remote sensing big data: A technical review. *Big Earth Data*, *6*(4), 527–560. https://doi.org/10.1080/20964471.2021.1964879

Zuo, C., Ding, L., Yang, Z., & Meng, L. (2022). Multiscale geovisual analysis of knowledge innovation patterns using big scholarly data. *Annals of GIS*, *28*(2), 197–212. https://doi.org/10.1080/19475683.2022.2027012

About the Author

Murat KILINÇ, is an Assistant Professor at Karadeniz Technical University. He received his Bachelor's degree in Mathematics and Computer Science from Eskişehir Osmangazi University in 2014, his Master's degree in Management Information Systems from Dokuz Eylül University in 2018, and his Ph.D. in the same field from the same university in 2022. Specializing in artificial intelligence, machine learning, and data analytics, his research focuses on machine learning applications in business intelligence, crowdfunding, and virtual reality technologies. He has also contributed to various collaborative projects aimed at enhancing accessibility and integrating artificial intelligence in education. Murat Kilinc's work aims to bridge the fields of business and computer science by developing interdisciplinary solutions that improve decision-making processes and optimize operations through advanced technology.
E-mail: muratkilinc@ktu.edu.tr, **ORCID:** 0000-0003-4092-5967

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 9%.

The New Role of Teachers in the Age of Web 3.0: Personalized Learning Environments through Learning Analytics

Hüseyin ULUKUZ

Ministry of National Education of Türkiye

To Cite This Chapter

Ulukuz, H. (2024). The New Role of Teachers in the Age of Web 3.0: Personalized Learning Environments through Learning Analytics. In M. Hanefi Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 20-27). ISRES Publishing.

Introduction

The first version of the web, Web 1.0, introduced in 1989, was characterized by a limited number of content creators and a wide readership, providing a platform for accessing information rather than contributing to it. This initial phase of the internet primarily offered a read-only experience, with websites designed for one-way communication and minimal user interaction (Dominic, Francis & Pilomenraj, 2014). In contrast, Web 2.0, emerging in 1999 and still widely used today, shifted the paradigm to enable users to actively create content, allowing numerous individuals to reach large audiences. Whereas Web 1.0 was primarily consumption-oriented, Web 2.0 is centered on participation and content generation. While Web 2.0 technologies have significantly enhanced user engagement and social inclusion, a paradigm shift in web technologies has redirected the focus from mere access to information towards a deeper emphasis on the meaning and personalization of information. This has led to the emergence of Web 3.0 technologies called the semantic web (Fırat & Fırat, 2021). The Semantic Web builds upon the foundational concepts of the World Wide Web, introducing "meaning" to web content in a way that allows machines to interpret the significance of information. (Berners-Lee et al., 2001). However, Web 3.0 is not solely defined by the semantic web; rather, it encompasses a broader spectrum of concepts and functionalities. Specifically, it is founded on core principles such as decentralization and interoperability, representing an idealized vision of the internet (HyScaler, 2024). Nevertheless, its full potential remains constrained by current technological limitations.

In this context, the concept of learning analytics—a rapidly growing field within instructional technology—plays a crucial role in establishing personalized learning environments and addressing various learning styles by tailoring educational responses to align with individual learning needs. This field leverages the power of big data through Web 3.0 technologies, such as "semantic web" or "smart web," to personalize and optimize educational processes (Merceron et al., 2015). Learning analytics is defined at the Learning Analytics Knowledge Conference as the collection, measurement, analysis, and reporting of data about learners and their contexts, with the aim of understanding and enhancing learning and the environments in which it occurs (Long & Siemens, 2011). This process is seen as a cyclical approach involving learners, data, metrics, and interventions (Clow, 2012). Consequently, a user-centered approach, in which teachers and students assume active roles, is crucial in the design of learning systems based on learning analytics (Quincey et al., 2019).

Taking all these developments into account, the literature on instructional technology

provides ample evidence that traditional learning environments have undergone substantial transformations, largely driven by advancements in web technologies (Brown & Green, 2018; Kimmons, 2020; Lai & Bower, 2020). Today, instructional technologies have reached a more advanced stage, largely due to the evolution of web technologies. In this sense, the ecosystem of Web 3.0, for instance, supports the integration of artificial intelligence and virtual reality into educational settings, providing significant contributions to learning environments (Johnson et al., 2011), thus allowing for immersive and adaptive educational experiences that cater to diverse learners. These developments have not only broadened access to educational resources but have also enabled more interactive and personalized learning experiences. As a result, educators are increasingly adopting online learning models, or at least blended approaches, that leverage these technologies to support diverse learning styles and enhance student engagement.

Traditionally, the role of teachers has been centered around the delivery of standardized content to a group of students, often through lectures and memorization-based instructional methods (Freire, 1970). In this framework, teachers act primarily as the central source of knowledge, dictating the pace and structure of the curriculum with limited adaptability to individual learning needs (International Society for Technology in Education, 2007). This approach generally assumes a one-size-fits-all model, where assessments are uniform and often serve more as a means of measuring retention rather than understanding. The emphasis is generally on conformity to predetermined learning outcomes, with less consideration for differentiated instruction or the unique context of each learner (Sullivan & Downey, 2015). As a result, opportunities for students to engage in critical thinking, creative problem-solving, or personalized inquiry can be limited, restricting tailored educational responses that foster a deeper comprehension of the material (Scott & Husain, 2021).

Statement of the Research Aim and Research Questions

This paper seeks to investigate whether the traditional role of the teacher can be redefined in learning environments, especially in contexts enhanced by learning analytics, where teachers tailor instructional methods to meet the specific needs of their students based on data insights. With the advent of learning analytics, teachers have access to a range of data that can provide insights into individual student performance and progress. By utilizing this data, teachers can gain a richer understanding of each student's strengths and weaknesses, enabling them to tailor their teaching methods and interventions accordingly. In addition to using learning analytics data for personalized instruction, it also has the potential to shift the focus from outcomes-based assessments to more process-oriented evaluations. This means that instead of solely measuring whether students have achieved predetermined learning outcomes, the emphasis is on evaluating how they are learning and engaging with the material. This shift towards a more process-oriented evaluation can enable teachers to identify areas where students may be struggling or thriving in their learning journey. In this context, exploring the potential impact of learning analytics on redefining the teacher's role in modern educational environments gains importance.

How Learning Analytics Supports Teachers in Learning Environments

Learning analytics consists of several key components that work together to improve educational experiences and outcomes. Fundamentally, it involves the collection, analysis, and interpretation of data, alongside the delivery of feedback and interventions aimed at learners and their learning environments (Long & Siemens, 2011). This process typically starts with data gathering, which can include information from learning management systems, assessment scores, and even student engagement metrics. Specifically, the collected data may encompass age, gender, socioeconomic status, learning disabilities or special needs, cultural background, attendance and participation records, exam/test results, project or group work performance, assignment submissions and grades, system login frequency, session duration, number of interactions, study hours and patterns, as

well as the courses and content accessed by students. By analyzing such data, educators can uncover patterns in student performance and behavior, allowing them to adapt instructional methods to address individual learning needs (Siemens, 2013).

The most visible application of learning analytics in online educational environments is through the use of dashboards, which can be categorized into two types: internal and external dashboards. Internal dashboards provide real-time feedback within the system, enabling both students and teachers to monitor progress and engagement instantaneously. External dashboards, on the other hand, function as reporting interfaces, offering detailed insights into the student's learning journey. These dashboards serve as tools for both students and educators to track performance, identify patterns, and support decision-making. Learning analytics can be further classified into four primary types: descriptive, diagnostic, predictive, and prescriptive analytics. Descriptive analytics gives teachers an overview of student performance by summarizing past data, helping them identify trends and patterns in learning behaviors. This information allows educators to adjust their instructional methods based on the overall performance of the class. In contrast, diagnostic analytics focuses on understanding why students succeed or struggle. By identifying gaps in knowledge or skills, teachers can design targeted interventions to address these issues effectively. Predictive analytics goes further by estimating the likelihood of student success and enabling educators to take proactive steps. It can help identify at-risk students early, enabling educators to provide timely support to those in need (Siemens, 2013). Finally, prescriptive analytics offers specific recommendations based on data insights, such as creating personalized learning pathways or suggesting resources to improve student engagement. Once students at risk of low performance are identified, prescriptive analytics provides actionable recommendations tailored to their specific needs. For example, it may suggest assigning additional math exercises to reinforce understanding or recommending one-on-one mentoring sessions to improve participation in live classes. Such recommendations are personalized to address individual student requirements, helping to achieve more effective learning outcomes. Together, these types of analytics provide a powerful framework that supports personalized learning and helps teachers improve their instructional practices, leading to better educational outcomes for all students. As educators embrace these tools, they not only enhance their effectiveness but also foster a collaborative atmosphere where students feel valued and understood. Thus, integrating learning analytics into educational practices is a crucial step toward fostering a more inclusive and adaptive educational system (Siemens & Baker, 2022).

Pedagogical Goals of Learning Analytics and Its Role in Learning

Learning analytics serves several pedagogical goals, primarily focusing on the personalization of educational experiences, supporting student self-regulation, and aiding teachers in planning and managing learning. To achieve these goals, data-driven methods are employed to adapt educational experiences to the unique needs of students. This process requires the systematic collection, reporting, analysis, and interpretation of data within the learning environment, generating meaningful insights that inform instructional decisions. Specifically, it involves creating profiles for students, which can be either static or dynamic, and adjusting educational strategies accordingly (Arnold et al., 2012). Furthermore, such methods can be applied to groups of learners, enabling collaborative and group-based learning. Thus, this approach not only enhances personalized learning but also fosters a collaborative atmosphere where students can benefit from shared knowledge and experiences. Techniques within learning analytics, such as predictive modeling, clustering to identify groups of students with similar profiles, and association rule mining, are commonly employed to facilitate both individual and group learning (Sclater, 2017). On the other hand, such approaches align with the constructivist theory, which emphasizes that students develop their understanding through active engagement, experiences, and reflection. In this context, learning analytics supports the constructivist

paradigm by providing real-time feedback and facilitating the development of personalized learning pathways that cater to individual student needs.

Another significant pedagogical use of learning analytics is to help students monitor and improve their own learning processes. This model is grounded in psychological theories such as experiential learning, self-reflection, and self-regulated learning (Huberth et al., 2015). Various tools, including internal and external dashboards supported by learning analytics in educational environments, provide students with feedback on their academic status and study habits. These dashboards can be integrated directly into the learning environment or accessed through external channels like email messages or real-time interfaces. This adaptive approach not only addresses the diverse needs of learners but also fosters a sense of ownership over their learning processes. For instance, when students receive timely feedback thanks to learning analytics tools, they can identify their strengths and areas for improvement more effectively, leading to better decision-making regarding their study habits and strategies. The feedback mechanism is particularly beneficial in online collaborative settings, where students may struggle with self-discipline and motivation. Despite these advancements, students may still struggle to interpret and utilize the data provided by such analytics-supported systems due to their limitations in understanding the feedback. This difficulty may arise from the complexity of data representation or the insufficient digital and data literacy skills of students. Therefore, Arnold et al. (2012) claim that the most effective systems integrate analytic feedback with structured support from teachers to ensure meaningful use of the information. Teachers, leveraging learning analytics data, can provide context and guidance, helping students to decode the feedback and apply it constructively. This support is crucial in bridging the gap between data interpretation and practical application. By offering personalized insights and actionable advice, educators can empower students to take ownership of their learning journey. Thus, the guiding role of an educator remains an indispensable element in systems supported by learning analytics. This ensures that students can make meaningful use of the data to enhance their learning outcomes, ultimately leading to improved academic performance and greater self-efficacy. Furthermore, this collaborative approach fosters a supportive learning environment where students feel more confident and motivated to engage with their studies.

Learning Environments and Power Dynamics in the Light of Learning Analytics

In traditional educational settings, educators hold absolute power and authority. However, in today's environments where digital technologies are widely utilized, the tools themselves, due to their inherent nature, tend to constrain or reshape the instructor's authoritative role. Canadian communication theorist Marshall McLuhan encapsulates this phenomenon in his famous phrase, "The medium is the message" (1967). In other words, the form of a message dictates how it is perceived, which may unintentionally shape or undermine the instructor's authority over the message or the educational process. To put it simply, an instructor becomes reliant on the technology they use, shaping their teaching within the capabilities and limitations defined by that technology. Furthermore, this dynamic may vary based on the degree to which the technology dominates the educational environment (Whitworth, 2014).

In the 21st century, advancements in digital technologies have enabled the collection of context-specific data from various platforms, paving the way for personalized guidance, manipulation, and information dissemination. Within the context of educational technologies, this data is made meaningful through learning analytics, a specialized application of data science tailored to educational environments. Learning analytics provides educators with valuable insights into their students' learning behaviors, preferences, and progress. While this meaningful data offers significant benefits, it also influences how educators interpret and assess their students. Since this guidance is data-driven, its accuracy is generally high. However, as the renowned Turkish intellectual

Cemil Meric aptly states, “Every definition is a distortion“ (1974). This reminds us that even the most accurate data interpretations are subject to the limitations and biases inherent in their framing. Therefore, educators must remain critical and reflective when using data to inform their teaching practices, ensuring that they consider the broader context and individual student needs.

Despite potential challenges, learning analytics significantly enhance educational environments by profiling students and equipping teachers with accurate and, at times, previously unrecognized insights. These insights facilitate a more personalized approach to addressing individual learning needs. From a technical standpoint, the categories of learning analytics (descriptive, diagnostic, predictive, and prescriptive) illustrate a progression wherein each successive level involves increasingly interpretive and inferential analyses of collected data, moving beyond basic description. Artificial Intelligence (AI) models play a critical role in the processes by interpreting complex data relationships, offering teachers an additional layer of insight or even a form of augmented consciousness. This aligns with Bakhtin’s (1984) concept of truth, which he defines as requiring a diversity of consciousnesses—truth that cannot reside within a single perspective but emerges from the interaction and collision of multiple consciousnesses, inherently “full of event potential“ (p. 81). In the context of this paper, Bakhtin’s notion of truth can be applied to the function of advanced learning analytics. These analytics provide data-driven insights that enable teachers to develop a deeper and more nuanced understanding of their students, allowing them to engage with students’ realities—such as their readiness levels and unique learning contexts—with greater precision and adaptability.

Discussion

The emergence of Web 3.0 technologies and the integration of learning analytics into educational settings necessitate a reevaluation of the traditional role of teachers. Historically, teachers held an authoritative position, acting as the primary providers of knowledge and controlling the pace and direction of the learning process (Freire, 1970). However, as digital technologies such as learning analytics become more prominent, the power dynamics in educational environments are evolving. In Web 3.0-supported classrooms, the authority once held exclusively by teachers is now shared with, and at times challenged by, the authority of data itself.

Learning analytics, as defined by the Learning Analytics Knowledge Conference (2011), involves the collection, measurement, and analysis of data to understand and enhance learning and the environments in which it occurs. This data provides valuable insights into student behavior, progress, and engagement, creating a form of “data-driven authority“ in educational decision-making. While this shift emphasizes the significance of meaningful and actionable data, it also highlights the necessity of the teacher’s role as a mediator. Teachers now act as facilitators and moderators, leveraging learning analytics to create personalized learning pathways that address individual student needs (Siemens & Long, 2011).

This transformation aligns closely with the constructivist theory, which emphasizes active engagement, experiences, and reflection as the cornerstones of meaningful learning. Within this paradigm, teachers are no longer perceived as mere content providers but as facilitators of knowledge construction, who guide and support students in their individual learning journeys. This shift reflects the growing recognition that traditional, teacher-centered models of education are insufficient to meet the diverse and evolving needs of students in the age of Web 3.0. Consequently, it is necessary to expect today’s teachers to adapt to more secondary and supportive roles that emphasize collaboration and flexibility.

By embracing these roles, teachers can align their practices with the demands of educational environments strongly supported by digital technologies. As Quincey et al.

(2019) highlight, this adaptation enables educators to design learning environments that integrate insights from learning analytics to foster collaboration, critical thinking, and creativity. Such environments not only respond to individual learning needs but also promote an inclusive and dynamic atmosphere where students can actively engage with the material and with one another. This redefined role of the teacher, supported by data-driven insights, underscores the importance of creating educational experiences that are both personalized and participatory, ensuring that students develop the skills and mindset required to navigate the complexities of the digital age.

However, this evolving role is not without its challenges. Data, despite its precision, is inherently shaped by biases and contextual limitations, requiring educators to approach it with a critical and reflective mindset. Teachers must critically interpret and contextualize data, ensuring that it complements rather than replaces their professional judgment (Whitworth, 2014). While data provides valuable insights, it cannot capture the relational and emotional dimensions of teaching. Teachers must balance the authority of data with their ability to inspire, connect with, and motivate students, ensuring that the humanistic elements of education are preserved.

Moreover, students often face difficulties in interpreting and utilizing the feedback provided by learning analytics-supported systems, due to their limited digital and data literacy (Arnold et al., 2012). This further emphasizes the indispensable role of teachers as guides who help students decode and apply these insights effectively. For instance, when real-time dashboards highlight areas for improvement, it is the teacher who provides the context and guidance necessary for students to act on this information. By integrating their expertise with data insights, teachers can empower students to take ownership of their learning processes, fostering a sense of self-regulation and autonomy.

In summary, the integration of Web 3.0 technologies and learning analytics is expected to lead to a transformation in the role of teachers. While their traditional authority may have diminished in some respects, their role has evolved into that of a data-informed facilitator and mentor. Teachers remain central to the learning process, bridging the gap between data insights and actionable strategies while fostering a collaborative and inclusive learning environment. This redefined role not only aligns with the demands of the digital age but also ensures that education remains a holistic, student-centered endeavor in an increasingly data-driven world. As Bakhtin (1984) suggests, truth emerges from the interaction of multiple perspectives, and in education, it is through the interplay of teacher expertise, student engagement, and data-driven insights that meaningful learning experiences are created.

Conclusion

The integration of Web 3.0 technologies and learning analytics is redefining the role of teachers, transforming them from authoritative content providers into facilitators and moderators in personalized learning environments. By leveraging data-driven insights, teachers can design targeted interventions, address individual learning gaps, and create more inclusive and adaptive educational experiences. However, this reliance on data also introduces challenges. The interpretation of data is shaped by biases and contextual limitations, requiring educators to critically assess analytics while maintaining the relational and humanistic aspects of teaching. Training in digital and data literacy is essential to help teachers effectively bridge the gap between technology and pedagogy. In the age of Web 3.0, teachers remain central to the learning process, contextualizing data to foster meaningful, student-centered education. By embracing these tools and practices, educators can navigate the challenges of the digital age while creating innovative and impactful learning environments that prepare students for a collaborative and dynamic future.

References

Arnold, K. E., Lynch, G., Huston, D., Wong, L., Jorn, L., & Olsen, C. W. (2014, March). Building institutional capacities and competencies for systemic learning analytics initiatives. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 257-260).

Bakhtin, M. (1984). *Problems of Dostoevsky's poetics* (Vol. 8). U of Minnesota Press.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 28-37.

Brown, A., & Green, T. (2018). Issues and trends in instructional technology: consistent growth in online learning, digital content, and the use of mobile technologies. *Educational Media and Technology Yearbook: Volume 41*, 61-71.

Clow, D. (2012, April). The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 134-138).

Dominic, M., Francis, S., & Pilomenraj, A. (2014). E-learning in web 3.0. *International Journal of Modern Education and Computer Science*, 6(2), 8.

Fırat, E. A., & Fırat, S. (2020). Web 3.0 in learning environments: A systematic review. Turkish Online Journal of Distance Education, 22(1), 148-169.

Freire, P. (2020). Pedagogy of the oppressed. In *Toward a sociology of education* (pp. 374-386). Routledge.

Huberth, M., Chen, P., Tritz, J., & McKay, T. A. (2015). Computer-tailored student support in introductory physics. *PloS one*, *10*(9), e0137001.

HyScaler. (2024, June 4). *The core principles of Web 3.0: Decentralization and interoperability*. HyScaler. https://www.hyscaler.com/web3-core-principles

International Society for Technology in Education. (2007). *National educational technology standards for students*. ISTE (Interntl Soc Tech Educ.)

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K., (2011). *The 2011 Horizon Report*. The New Media Consortium.

Kimmons, R. (2020). Current trends (and missing links) in educational technology research and practice. *TechTrends*, *64*(6), 803-809.

Lai, J. W. M., & Bower, M. (2020). Evaluation of technology use in education: Findings from a critical analysis of systematic literature reviews. *Journal of Computer Assisted Learning, 36*(3), 241-259.

Long, P., & Siemens, G. (2011). LAK'11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada. ACM, New York, NY.

McLuhan, M. (1967). The medium is the massage. *A Benthm Bbook*.

Merceron, A., Blikstein, P., & Siemens, G. (2015). Learning analytics: From big data to meaningful data. Journal of Learning Analytics, 2(3), 4-8.

Meriç, C. (1974). *Bu Ülke*. İstanbul: İletişim Yayınları.

Quincey, E., Briggs, C., Kyriacou, T., & Waller, R. (2019, March). Student centered design of a learning analytics system. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 353-362).

Sclater, N. (2017). *Learning analytics explained*. Routledge.

Scott, T., & Husain, F. N. (2021). Textbook reliance: Traditional curriculum dependence is symptomatic of a larger educational problem. *Journal of Educational Issues*,

7(1), 233-248.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380-1400.

Sullivan, S. C., & Downey, J. A. (2015). Shifting educational paradigms: From traditional to competency-based education for diverse learners. *American Secondary Education*, 4-19.

Whitworth, A. (2014). *Radical information literacy: Reclaiming the political heart of the IL movement*. Elsevier.

About the Author

Dr. Eng. Hüseyin ULUKUZ, is an Education Specialist at the Republic of Türkiye, Ministry of National Education, General Directorate of Innovation and Educational Technologies, where he has been serving for the past year. He holds a Doctorate in Educational Technology from the University of Manchester, with a dissertation titled: *Crafting Participatory Experiences in the Web 2.0 Era.* His areas of expertise include Immersive Technologies, Web 2.0 Applications, and Learning Analytics. Dr. Ulukuz's research and professional focus are centered on integrating innovative digital tools into educational settings to enhance engagement, participation, and personalized learning experiences.

E-mail: huseyin.ulukuz@hotmail.com, **ORCID:** 0000-0003-2272-6112

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 6%.

Pioneering Data-Driven Decisions: The Future of Predictive Modeling in Data Science

Kartal DERİN

Smartpro Technology

To Cite This Chapter

Derin, K. (2024). Pioneering Data-Driven Decisions: The Future of Predictive Modeling in Data Science. In M. Hanefi Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 28-37). ISRES Publishing

Introduction

In today's rapidly evolving digital era, businesses across numerous sectors are increasingly turning to data-driven decision-making to enhance operational efficiency and achieve strategic objectives. At the core of this shift lies predictive modeling, a powerful analytical tool that leverages historical data to anticipate future outcomes. By employing machine learning algorithms and statistical techniques, predictive models continuously improve, providing organizations with unparalleled insights into consumer behavior, market trends, and operational risks. This paper examines the profound impact of predictive modeling across various industries, including healthcare, education, business, and technology. From AI-driven diagnostic tools to real- time gesture recognition systems and drones, predictive analytics is driving innovation in diverse fields. The study also discusses the evolution of predictive modeling, tracing its roots from traditional statistical methods to advanced machine learning techniques that form the foundation of future data science.

As emerging technologies like quantum computing, federated learning, and edge AI come into play, predictive modeling is poised to have an even greater impact. This paper offers a comprehensive overview of the current state of predictive modeling and its future potential, including an analysis of its applications and ethical considerations.

Figure 1

Predictive Modeling (original)

The Development and Progression of Predictive Modeling

Predictive modeling began with traditional statistical approaches, which offered a foundation for identifying patterns in data and predicting future events. Early methods, such as regression analysis and decision trees, were effective for smaller, less complex datasets. However, as computational power grew and data collection methods advanced, the demand for more sophisticated models emerged, leading to significant breakthroughs in the field.

By the late 20th century, machine learning began to replace traditional rule-based systems with more flexible models. Algorithms like support vector machines, decision trees, and neural networks enhanced the accuracy and complexity of predictive modeling. These algorithms could learn from vast datasets and uncover intricate patterns that were previously undetectable. The combination of big data, improved computational power, and advanced algorithms helped predictive modeling become widely used across industries during the early 21st century.

In recent times, predictive modeling has integrated deep learning techniques, allowing models to autonomously enhance their performance. Deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have enabled predictive analytics to excel in tasks such as image recognition, natural language processing, and behavioral prediction.

Figure 2

Machine Learning and Deep Learning (original)

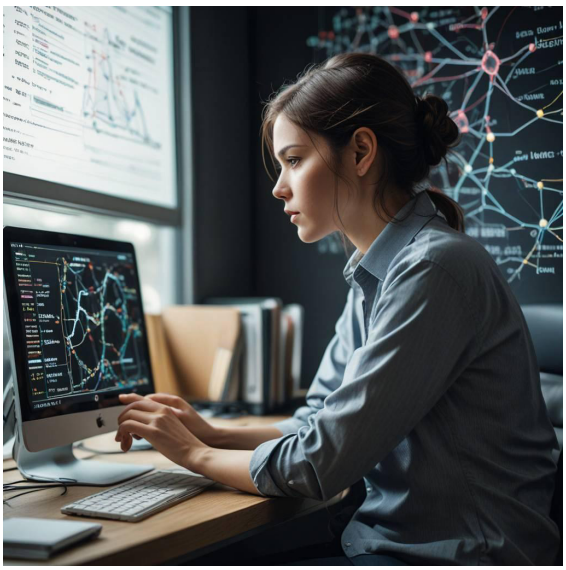
Machine Learning and Deep Learning in Predictive Modeling

The widespread application of machine learning (ML) and deep learning (DL) has fundamentally reshaped the landscape of predictive modeling. Where traditional models faced challenges with smaller datasets, ML algorithms now allow for the sophisticated analysis of much larger and more complex datasets. The ability of ML to refine models and improve prediction accuracy has introduced significant advancements in predictive analytics.

Machine learning uses both supervised and unsupervised techniques to reveal hidden patterns within data, leading to more accurate forecasts. Foundational ML methods, such as regression, decision trees, and support vector machines, have proven successful on large-scale datasets (Ghosh, 2021). Brown and Smith (2018) highlight that advanced machine learning techniques have significantly improved prediction accuracy, especially for large and complex datasets. On the other hand, deep learning is especially useful for

tackling unstructured or highly intricate data.
Deep learning models are inspired by neural architectures akin to the human brain, using layers of neural networks to process complex datasets. These models excel in areas like image recognition, natural language processing, and time-series analysis. Convolutional neural networks (CNNs) are frequently employed for tasks involving image recognition, while recurrent neural networks (RNNs) are better suited for sequential data processing.

Figure 3
Challenges and Opportunities (original



Data Sources for Predictive Modeling: Challenges and Opportunities

As predictive modeling advances, the diversity and complexity of data sources have significantly expanded. Traditionally, models relied on structured data from databases and historical records. However, with the advent of the Internet of Things (IoT), wireless sensing technologies, and social media, new dynamic data streams have emerged. While these sources present opportunities for deeper insights, they also introduce challenges related to data quality, consistency, and real-time processing.
The Internet of Things (IoT) has become a significant contributor to the vast amount of data available for predictive modeling. IoT-connected devices generate data that can predict everything from equipment malfunctions in industrial settings to consumer behaviors in smart homes.

However, IoT data often comes in unstructured or semi-structured forms, requiring advanced preprocessing to be used effectively.

Wireless sensing technologies, such as Channel State Information (CSI), add another dimension to predictive modeling. CSI data captures real-time wireless signals, which can be leveraged for human behavior detection systems, including gesture recognition. However, due to the large volume and variability of this data, sophisticated algorithms are needed to filter out noise and detect relevant patterns .

Figure 4
Human Gesture Recognition (original)

The Role of Neural Networks in Human Gesture Recognition

Neural networks have dramatically transformed human gesture recognition, providing powerful tools to interpret and predict human movements with remarkable accuracy. By leveraging deep learning architectures like CNNs and RNNs, gesture recognition systems can process and analyze real-time data to decode complex hand movements, body gestures, and facial expressions. These systems are applied in a variety of fields, ranging from human-computer interaction to advanced robotics and healthcare.

Figure 5
Real-World Applications (original)

Real-World Applications: Predictive Modeling in Industry

Predictive modeling has reshaped various industries by streamlining processes and enhancing decision-making. By utilizing large datasets with advanced algorithms, industries such as finance, manufacturing, supply chain management, and aerospace have significantly improved their performance, reduced costs, and forecasted future trends

more accurately. The ability to offer actionable insights has made predictive modeling essential in today's highly competitive market landscape.

One notable application is financial forecasting. Financial institutions utilize machine learning models to analyze historical data, market trends, and economic indicators to predict future market movements. These models allow businesses to mitigate risks by detecting early signs of economic downturns, stock price fluctuations, and potential credit defaults. Additionally, banks rely on predictive models to assess the creditworthiness of their customers, offering more personalized and accurate loan offerings.

In supply chain management, predictive modeling helps companies optimize their operations by forecasting demand, managing inventory, and identifying potential disruptions. For instance, predictive models can anticipate changes in product demand by analyzing historical sales data, seasonal patterns, and real-time market fluctuations, allowing businesses to adjust production schedules and resource allocation efficiently.

Predictive modeling is also invaluable in project management, where it is used to forecast project timelines, predict budget overruns, and estimate resource requirements. By examining historical project data, managers can anticipate potential risks and inefficiencies, leading to more precise resource planning and timely project completion. Predictive models are particularly useful in large- scale software development projects, where delays can be costly and client satisfaction is paramount.

Figure 6

Real-World Applications2 (original)

In the aerospace industry, predictive modeling has become crucial for optimizing maintenance and improving safety protocols. Airlines rely on predictive maintenance models to forecast potential equipment failures, thereby minimizing costly downtime and ensuring passenger safety. By analyzing data from aircraft sensors and maintenance logs, these models predict when components are likely to fail, allowing preventive maintenance to be scheduled before an issue becomes critical.

Predictive modeling also plays a role in AI-powered drones. These systems use machine learning to predict flight paths, optimize navigation, and anticipate environmental conditions such as wind speed or obstacles. In sectors like agriculture, construction, and logistics, AI-powered drones equipped with predictive models survey large areas, monitor crop health, and deliver goods with precision. These drones can operate autonomously,

adapting to dynamic environments in real-time (Zhang et al., 2021).

Finally, predictive modeling has found a place in the racing drone industry, where real-time data like speed, altitude, and trajectory are analyzed to optimize drone performance during races.

Predictive analytics helps identify potential risks or mechanical failures before they affect race outcomes, enhancing both performance and safety.

Figure 7
Healthcare and Education (original)

Predictive Analytics in Healthcare and Education

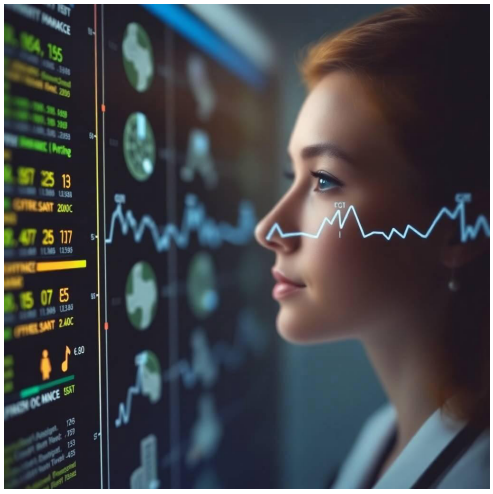
Predictive analytics has revolutionized healthcare and education by improving how these sectors utilize data to make informed decisions, enhance outcomes, and streamline operations. In healthcare, predictive models are used to diagnose diseases, forecast patient outcomes, and create personalized treatment plans. In education, predictive analytics powers personalized learning platforms, helping educators identify student needs and forecast academic performance.

In healthcare, predictive analytics plays a key role in the early detection of diseases. By analyzing large datasets, including medical histories and imaging results, predictive models can identify patterns that signal the onset of diseases. For example, these models are increasingly being used to detect cancer at earlier stages by analyzing radiological images. Machine learning algorithms pick up on subtle changes in tissue that may not be visible to the naked eye, leading to earlier and more accurate diagnoses.

Additionally, predictive models can help anticipate outcomes for chronic conditions such as diabetes or heart disease. Data from wearable devices and real-time monitoring systems allow healthcare providers to intervene early, reducing hospitalizations. Predictive analytics also improves healthcare efficiency by predicting patient admissions and helping hospitals allocate resources better during peak periods .

In education, predictive analytics is transforming traditional models by providing personalized learning experiences. By analyzing student performance and engagement data, predictive models identify areas where students may need additional support, allowing teachers to tailor their approach. This personalized method improves learning outcomes. Predictive models also assist institutions in identifying at-risk students, enabling timely interventions to increase retention (Johnson & Johnston, 2019).

Figure 8
Real-Time Decision-Making (original)



Real-Time Decision-Making: Predictive Models in Action

One of the most valuable outcomes of advances in machine learning and AI is real-time decision- making, made possible by predictive models. Real-time analytics enables industries to respond quickly to changing environments, thereby improving decision-making efficiency and accuracy. In manufacturing, real-time object-counting systems powered by predictive models detect and track items on conveyor belts, ensuring precise counting and quality control. For example, in the food processing industry, predictive models count items such as oranges in real time, optimizing operations and reducing errors (Gonzalez & Peters, 2021). Similarly, in human gesture recognition, systems that integrate machine learning algorithms with computer vision allow users to control devices through hand or body movements. These systems are increasingly used in consumer electronics and healthcare for contactless interfaces. Real-time predictive models are also essential in healthcare for monitoring critical patients, allowing providers to intervene when necessary .

In autonomous systems, such as drones and self-driving cars, predictive models process real-time sensor data to navigate environments, avoid obstacles, and make quick decisions. This increases safety and operational efficiency, allowing drones to adjust to changing conditions such as weather or terrain (Zhang et al., 2020).

Figure 9
Emerging Trends (original)

The Future of Predictive Modeling: Emerging Trends

The future of predictive modeling is being driven by several important trends, including quantum computing, federated learning, and edge AI. These emerging technologies are enhancing the speed, precision, and scalability of predictive models, making them more effective and accessible for various industries.

Quantum computing holds significant potential to transform predictive modeling by processing enormous amounts of data at unprecedented speeds. In contrast to classical computers, which use binary bits, quantum systems rely on qubits, which can handle more complex calculations in less time. This improvement is especially advantageous for predictive analytics, where large data sets with numerous variables need to be processed efficiently. Quantum computing is expected to significantly advance machine learning algorithms, unlocking solutions to previously unsolvable challenges.

Federated learning is another innovative approach, addressing privacy issues by training models on decentralized data sources. This technique is particularly valuable in fields like healthcare and finance, where regulatory compliance and privacy are paramount. Federated learning allows predictive models to harness diverse data sets while safeguarding sensitive information.

Edge AI, which involves deploying predictive models directly on devices such as smartphones and sensors, eliminates the need for cloud-based processing. This minimizes latency and supports real- time decision-making, which is essential for applications such as self-driving vehicles. As edge computing technologies progress, predictive models will be able to process data locally, boosting the speed and efficiency of various operations .

Finally, Explainable AI (XAI) is gaining momentum as it addresses the issue of transparency in machine learning models. XAI aims to make models more understandable, providing insights into decision-making processes. This is crucial for industries like healthcare and finance, where transparency in decision-making builds trust and accountability (Miller, 2019).

Figure 10

Responsible AI(original)

Ethical Considerations and Responsible AI in Predictive Modeling

As predictive modeling becomes an essential part of decision-making across various industries, it raises critical ethical concerns like fairness, transparency, bias, and

accountability. While predictive models provide significant insights, they also present challenges, particularly in maintaining fairness.

One of the primary concerns is bias. Predictive models, when built on biased data, may inadvertently reinforce or even magnify existing inequalities, resulting in unfair outcomes. This issue has been particularly highlighted in algorithms used in areas such as criminal justice and recruitment, where minority groups may be disproportionately disadvantaged. To combat this, it is crucial to utilize fairness-aware algorithms and ensure the inclusion of diverse, representative datasets that accurately reflect the population being analyzed.

Transparency is another major issue, especially with complex machine learning models that often function as "black boxes," making it difficult to interpret how decisions are made. Garcia and Lopez (2020) emphasize that addressing ethical challenges, such as transparency and bias, is crucial for building trust and ensuring fair outcomes in AI-driven predictive analytics. In sensitive areas like healthcare, it's vital for stakeholders to understand the rationale behind model decisions to foster trust. Explainable AI (XAI) provides a solution by clarifying how models arrive at their conclusions, thereby improving both accountability and trustworthiness.

Accountability is equally important as predictive models become embedded in the decision-making frameworks of organizations. Establishing proper oversight and audit mechanisms ensures that the ethical implications of these models are continuously evaluated. Regular assessments can help guarantee that predictive models remain fair and accurate.

Finally, privacy concerns become more pressing as predictive models increasingly rely on personal data. Federated learning offers a promising approach by allowing decentralized data processing, which helps safeguard privacy while still enabling robust predictive analysis.

Figure 12
Ethical (original)

Conclusion

Predictive modeling has firmly established itself as a key tool in modern data science, empowering industries to make informed, data-driven decisions. Technological advancements in machine learning, deep learning, and artificial intelligence have brought transformative improvements across fields like healthcare, finance, education,

and manufacturing, enhancing operational efficiency, reducing risks, and optimizing decision-making processes.

The adoption of emerging technologies like quantum computing, federated learning, and edge AI will further boost the capabilities of predictive models, making them faster, more scalable, and more accurate. These advancements will enable predictive analytics to expand into new industries while addressing ethical issues like bias, transparency, and accountability.

To support the continued development of predictive modeling, it's essential to prioritize responsible and ethical usage. Focusing on fairness, transparency, and accountability will allow organizations to maximize the advantages of predictive models while minimizing potential risks, ensuring that these tools benefit all stakeholders.

References

Brown, J., & Smith, L. (2018). Advanced machine learning techniques in predictive modeling. Journal of Data Science and Applications, 6(4), 340–356. https://doi.org/10.1016/j.jdsapp.2018.06.003

Chen, P., & Zhang, Z. (2014).Title: Big data applications, opportunities, and challenges. Journal: Artificial Intelligence Review.DOI: https://doi.org/10.1007/s10462-014-9445-3

Garcia, M., & Lopez, A. (2020). Ethical challenges in AI-driven predictive analytics. AI Ethics and Society, 12(1), 45–60. https://doi.org/10.1080/AI.2020.0001

Montanaro, A. (2021).Title: Quantum computing and its impact on predictive analytics. Journal: Computing Futures, 5(3),67–79.DOI:https://doi.org/10.1016/j.cofut.2020.100051

Toma, M., & Ong Chi Wei, O. (2023).Title: Predictive modeling in medicine: A review. Journal: MDPI Journal of Medicine, 3(2), 590–601. DOI:https://doi.org/10.3390/encyclopedia3020042

About the Author

Kartal DERİN is a Data Scientist and the Head of the AI Department at SmartPro Teknoloji in İstanbul. With certifications from IBM, HP, and Google in data science and artificial intelligence, his expertise includes artificial intelligence, machine learning, and deep learning applications. Kartal has led and contributed to numerous global projects, particularly in healthcare and drone technology. His proficiency in Python programming has allowed him to develop advanced AI-powered solutions across various sectors. He is an active participant in AI research and frequently shares his insights and innovations on LinkedIn.

Email:kartalderinmail@yahoo.com ,**ORCID:** 0009-0009-7104-5955

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 6%.

An Overview of Social Network Analysis: Metrics, Tools and Applications

Akça Okan YÜKSEL

Middle East Technical University

To Cite This Chapter

Yüksel, A. O. (2024). An Overview of Social Network Analysis: Metrics, Tools and Applications. In M. Hanefi Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 38-56). ISRES Publishing

Introduction

A social network is defined as a collection of social entities, such as individuals, groups, and organizations, connected by relational data with some interactions or relationships between them. Examples of such networks include friendship networks, follower networks, interaction networks, co-authorship networks, and spread networks (Tabassum et al., 2018). The two main components of any social network are entities and relationships (Scott, 2000). The combination of these two elements creates a social network. Entities may be individual people or collective actors, such as groups and organizations. Common examples of individual actors include students in a school, employees in a corporate firm, or members of a political organization. Collective actors could be companies, foundations, or political parties. Sometimes, networks consist of different types of entities, such as a healthcare system or an education system. A relationship is generally defined as a specific type of contact, connection, or bond between a pair of entities or a dyad (Wasserman & Faust, 2004). Relationships can be directed, where one actor initiates and the other receives (e.g., giving advice, selling), or undirected, where reciprocity occurs (e.g., chatting, collaborating). A relationship is not a characteristic of a single entity but is a common dyadic property that exists as long as both participants maintain it. The diverse relationships between individual and collective entities can represent network structures and explain their impacts. The specific type of relationship a researcher should measure depends on the research objectives. For example, a study on community networks will likely examine various neighborhood activities, whereas a study on banking networks will focus on financial transactions. Borgatti et al. (2009) classified the types of relationships in social networks. These classifications and examples related to them are presented in Table 1.

Table 1
Types of Relationships

Relationship Type	Example
Similarities	Being on the same team, attending the same school, same gender, similar hobbies
Relationships	Kinship, marriage, friendship

Interactions	Help, advice, recommendation
Flows	Information flow, personnel changes, international trade

Social networks are suitable subjects for both quantitative and qualitative studies as they contain both the structure and the content of social relationships (Coviello, 2005). The analytical approach used in studying social networks is Social Network Analysis. Social Network Analysis (SNA) is a powerful analytical method aimed at examining connections and interactions between individuals, groups, institutions, or devices to make inferences from these relationships (Edwards, 2010). SNA finds application not only in personal relationships and social circles but also in business, healthcare, education, biology, and many other fields. This analytical method provides data scientists, researchers, and analysts with a broad and versatile toolkit, enabling the understanding of complex network structures and the extraction of information from them.

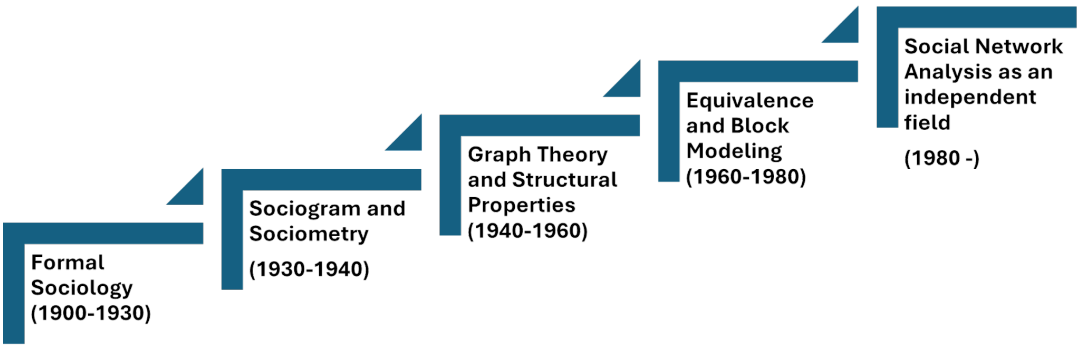
The rapidly growing subset of social networks is social media. Applications like Instagram, Twitter, LinkedIn, and WeChat facilitate daily information exchange. A significant concern is how the vast, complex data generated by online social network users can be searched, retrieved, stored, shared, processed, and visualized. SNA has become a widely applied method in research to investigate networks of relationships at individual, organizational, and societal levels. With the popularization of social networking sites like Facebook, Twitter, and Instagram, and the development of automated data collection techniques, the demand for SNA has recently increased significantly.

Social network analysis, with its interdisciplinary approach, is used in various fields and is of considerable importance. It offers the opportunity to analyze the relationships between social entities and the significance of these relationships (Oliveira & Gama, 2012). It allows the identification of similarities and differences in relationships between entities within social networks (Somyürek & Güyer, 2020). It also enables the integration of relationships and attributes in data structures (DeJordy & Halgin, 2008). In social network analysis, the visual presentation of data allows the entire structure to be seen as a whole and provides insight into the dynamics and effectiveness of any network (Lewis, 2011).

Historical Development of Social Network Analysis (SNA)

Social network analysis is a method that has evolved over time with the integration of various disciplines and has become an important tool in modern social sciences today. Social network analysis (SNA) has been used for a long time to represent complex relationships between participants in social systems at all scales. It is summarized the historical development process of SNA in five main stages (Somyürek & Güyer, 2020) (Figure 1).

Figure 1
Historical Development of SNA



Formal Sociology (1900-1930)

- George Simmel laid the theoretical foundation of modern social network analysis by focusing on the formal analysis of social interactions.
- Simmel examined different patterns of relationships, such as dyadic and triadic interactions, and argued that sociology should focus on the forms of social relationships.

Sociogram and Sociometry (1930-1940)

- Jacob Levy Moreno developed the techniques of sociogram and sociometry to study interpersonal relationships in small groups.
- Sociograms provided a visual dimension to network analysis by graphically representing social ties within groups.

Graph Theory and Development of Structural Features (1940-1960)

- Heider's (1946) work on group dynamics and balance made significant contributions to SNA.
- Cartwright and Harary (1956) mathematically analyzed social relationships in terms of balance and group dynamics.
- Graph theory, with its structures of nodes and edges, enabled a better understanding of social relationships.
- Modeling positive and negative relationships allowed for the identification of structural features like density and clustering in social networks.

Equivalence and Block Modeling (1960-1980)

- Harrison White and his students defined social structures through roles and relationships.
- Granovetter (1983) demonstrated that weak ties could be more effective than strong ties, highlighting the functionality of social relationship networks.
- Block modeling was used to group structurally similar nodes and to determine the fundamental features of the network.

Social Network Analysis as an Independent Field (1980-Present)

- In the 1980s, SNA emerged as a distinct research field within social sciences.
- INSNA (International Network for Social Network Analysis) was established, the Sunbelt Conference was organized, and the journal *Social Networks* began publication.

SNA has since taken a more analytical approach, developing its methodologies, theoretical expressions, and software. Software like UCINET, Gephi, PAJEK, and R packages has been developed, broadening the application of SNA.

Graph and Metrics

Social networks are generally presented in 2 ways. The first one is graphs. Graphs are structures that visually share information about social networks. Mathematical operations cannot be processed indirectly with analysis. The other is matrices. Since matrices allow for computational operations, detailed information is shared by conducting butchered analyses (Streeter & Gillespie, 1992).

Social networks are typologically classified as directed-undirected and binary-valued. If there are arrows between the links in the representation of a network, it is defined as a directed network; if there are no arrows, it is defined as a non-directional social network

(Tunalı, 2016). The other is classified as valued and binary according to the value of the link. The first one is the type that expresses the presence or absence of the links between nodes as 0-1 expressed as binary. The other is the type where the numerical value of a link indicates the density, strength, frequency, or volume of connections between pairs of nodes. This type is called valued (Tunalı, 2016).

Centrality

Centrality is an essential metric that indicates which node has a critical position within the network. If an actor has a high centrality value, it shows that this actor holds a central position in the network (Bloch, Jackson & Tabaldi, 2023).

In calculating centrality, the nature of the relationship is considered. Centrality is calculated based on whether the relationship is directed, undirected, weighted, or unweighted. In undirected networks, the degree of a node is the number of connections that node has. In directed networks, incoming connections to the node are referred to as in-degree centrality, while outgoing connections are referred to as out-degree centrality. The sum of the in-degree and out-degree is the total degree of that node.

Degree Centrality

In a network graph, degree centrality is measured by the total amount of direct connections to other nodes. It indicates the level of outward connectivity. Higher values suggest greater connectivity, indicating how central the node is relative to other nodes in the network (Laghridat & Essalih, 2023). In-degree centrality is based on relationships initiated by other users towards a user, while out-degree centrality is based on relationships initiated by a user towards others. Degree centrality assumes that all neighbors in the network are equally important. What matters is the number of neighbors. However, in many cases, if a node is connected to powerful nodes, its importance increases.

For undirected networks:

- Degree Centrality = Node's degree (number of connections) / (N-1)

For directed networks, it is divided into in-degree and out-degree centrality:

- In-degree Centrality = Number of incoming connections to the node / (N-1)
- Out-degree Centrality = Number of outgoing connections from the node / (N-1)

Betweenness Centrality

Betweenness centrality is used to measure a node that plays an 'intermediary' role in a network (Marin & Wellman, 2011). If a node is located on the only path that other nodes need to traverse, such as communication, connection, transportation, or transaction, then this node must be important and is likely to have a high betweenness centrality.

The betweenness centrality CB(v) for a node in a non-directional network is calculated by the formula.

$$c_{betweeness}(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$$c_{betweeness}(v) = \frac{c_{betweeness}(v)}{(N-1)x(N-2)/2}$$

- s and t: Other nodes (node pairs) in the network,

- σ_{st} : The total number of shortest paths between nodes s and t,
- $\sigma_{st}(i)$: The number of shortest paths between nodes s and t that pass through node i
- N: The total number of nodes in the network

Closeness centrality

Closeness centrality is a measure of the total distance of a node from other nodes, if the length of the shortest paths of node N to other nodes in the network is small, then node N has a high closeness centrality (Zamanitajeddin et al., 2024). It refers to the convenience and ease of connections between the node of focus and other nodes.

Closeness centrality is calculated according to the following formula.
N: Total number of nodes in the network,
d (v,u): The length of the shortest path between node v and the other node u.

$$c_{closeness}(v) = \frac{(N-1)}{\sum_{u \neq v} d(v,u)}$$

The normalized closeness centrality for an undirected network can be expressed as follows.

$$c'_{closeness}(v) = \frac{c_{closeness}(v)}{(N-1)}$$

Eigenvector Centrality

This metric is based on assigning a relative score to each node and measures how well-connected a given actor is with other well-connected actors (Codal & Coskun, 2016). The main focus of eigenvector centrality is that the power and status of an actor is recursively defined by the power and status of its alters. Alters is a term often used in the egocentric approach of social network analysis and refers to actors who are directly connected to a particular actor, called the ego. What is noteworthy in this centrality measure is that the centrality of an individual depends on the centrality of all its neighbors with a positive constant. An individual with a high centrality of neighbors will also have a high centrality.

Eigenvector centrality is a more detailed version of degree centrality. It assumes that not all links have the same importance, taking into account not only the quantity but especially the quality of these links.

The eigenvector centrality x_i for node i in a network is calculated by the following formula.

Eigenvector centrality: $x_i = \frac{1}{\lambda} \sum_{j \in N(i)} x_j$

x_i: Eigenvector centrality for node i.

λ: A constant scaling factor, eigenvalue

N(i): Neighbors of node i

x_j: The centrality of node j

This formula states that the centrality of each node is a function of the centralities of

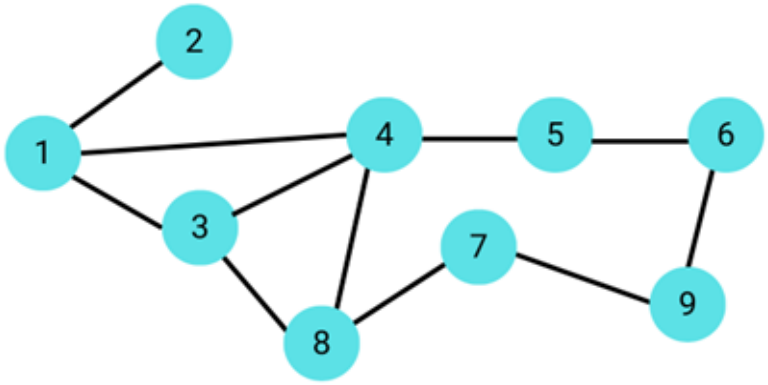
its neighbors. The eigenvalue λ is usually chosen to be the largest eigenvalue and the eigenvector centrality values are solved to find the eigenvector corresponding to the largest eigenvalue of the network.

Pagerank Centrality

PageRank centrality is a special type of eigenvector centrality and is the ranking criterion of the popular search engine Google (Tunalı, 2016). The three different factors that determine the PageRank of a node are the number of incoming links, the link propensity of anchors, and the centrality of anchors (Gençer, 2023).

An Example of a Social Network on Centrality Values

Figure 2
A Sample Social Network Graph



The edges representing the connections between nodes in the social network shown are listed as follows.

Edges: (1, 2); (1, 3); (1, 4); (3, 4); (3, 8); (4, 5); (4, 8); (5, 6); (6, 9); (7, 8); (7, 9)

Average Degree of the Network

First, it is calculated the degrees of the nodes in this network to calculate the average degree.

Degree of node 1: 3; Degree of node 2: 1; Degree of node 3: 3; Degree of node 4: 4; Degree of node 5: 2; Degree of node 6: 2; Degree of node 7: 3; Degree of node 8: 2; Degree of node 9: 2

Total degree of the nodes: 3 + 1 + 3 + 4 + 2 + 2 + 3 + 2 + 2 = 22

Since the total number of nodes is 9, the average degree is: 22 / 9 = 2.44

Table 2
Centrality Calculations for the Social Network in Figure 1

Node/ Actor	Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality	PageRank Centrality
1	0.375	0.471	0.250	0.408	0.135
2	0.125	0.333	0.000	0.143	0.055
3	0.375	0.533	0.095	0.480	0.127
4	**0.500**	**0.615**	**0.405**	**0.540**	**0.167**
5	0.250	0.500	0.202	0.231	0.094

6	0.250	0.421	0.095	0.118	0.099
7	0.250	0.444	0.155	0.185	0.095
8	0.375	0.533	0.262	0.422	0.129
9	0.250	0.381	0.071	0.106	0.099

Table 2 shows centrality calculations regarding Figure 1.

In terms of degree centrality, node 4 is in the most central position (0.500).

The node with the highest closeness centrality is node 4 (0.615), meaning that node 4 can reach other nodes in the network at shorter distances. Node A is positioned closest to the center, allowing it to connect with other nodes either directly or through short paths, which indicates that node 4 plays a key role in the information flow within the network. The node with the lowest closeness centrality is node 2 (0.000). This node is located at the most peripheral position in the network and must take longer paths to reach other nodes. The low closeness centrality of node 2 suggests that it is distant from the center of the network and participates less in information flow.

Betweenness centrality shows how much a node acts as a "bridge" between other nodes in the network. Calculations reveal that node 4 has the highest betweenness centrality value, at 0.405. According to these results, node 4 stands out as the most critical node on the shortest paths in the network. Node 2 has a betweenness centrality of zero, indicating that it does not act as a bridge between other nodes.

Eigenvector centrality is a measure based on how central a node's neighbors are. A node with high eigenvector centrality is more connected to central neighbors. Node 4, with the highest value, stands out as the most central node in the network, indicating that it is associated with neighbors who also hold highly central positions.

PageRank centrality calculates the importance of a node based on the centralities of the nodes that link to it. Originally used to rank web pages, this method is also widely used to identify influential nodes in networks. Node 4 has the highest PageRank value, indicating that it is in an important central position in the network. Node 2 has the lowest PageRank value, which suggests that its influence within the network is relatively low.

Local Clustering Coefficient

It is the ratio of the number of links between a node and its neighbors to the number of possible links they could have. In other words, it is a measure of the degree to which a node clusters with its neighbors.

The Local Clustering Coefficient indicates the probability that a node's neighbors will connect with each other. The value ranges from 0 to 1, with 1 indicating that all neighbors are connected.

The clustering coefficient C_i for a node i is calculated by the formula.

$$c_{cc(i)} = \frac{2x(numberoftriangles)}{dx(d-1)}$$

C_{CC} (i): local clustering coefficient for node i

d= Number of neighbors of the node

triangle= Number of available triangles

The local clustering coefficient of each node in an example social network of Figure 1 is calculated and presented in Table 3.

Table 3
Local Clustering Coefficient

Node	1	2	3	4	5	6	7	8	9
Local Clustering Coefficient	0.33	0.00	0.67	0.33	0.00	0.00	0.00	0.33	0.00

Node 3 has the highest clustering coefficient with a value of 0.667, indicating that its neighbors have strong ties to each other.

For nodes 2, 5, 6, 7, and 9, the clustering coefficient is 0, meaning that there are no connections between the neighbors of these nodes.

Whole Network Metrics

Size

The size of a network is determined by the number of actors in that network. When we consider a school as a social network, if both students and teachers play the role of actors, then all individuals will constitute the size of the network. When considering politics, the network size represents the number of people with whom an individual discusses political topics.

Average Degree

One of the metrics that defines the overall structure of a network based on degree is the average degree of all nodes in the network. This is calculated differently depending on whether the network is directed or undirected. Letting k_i be the degree of node i, the average degree of an undirected network with N nodes and E edges is calculated as follows:

$$\text{Average Degree} = \frac{1}{N}\sum_i^n k_i = \frac{2E}{N}$$

Density

The density of a network is equal to the total number of connections divided by the number of possible connections. The number of possible connections assumes that each person can have a connection with every other person. The normalized range varies from 0-1. It represents the extent of communication within the network. Higher numbers (above .03) indicate faster information diffusion and greater group cohesion (Aboelela et al., 2007).

The density of the network shown in Figure 1 is calculated as the number of available edges divided by the maximum number of possible edges.

The density of the network according to the formula density = $\frac{2xE}{Nx(N-1)}$

E= Number of edges; N = Number of nodes

In this network

Total number of nodes N=9

Number of sides E=11

The density is calculated as = (2x11) / 9 x (9-1) = 22 / 72 = 0.306. The density of this network corresponds to approximately 31%.

Centralization

Centralization is based on the extent to which the majority of links are connected to a small set of nodes (Scott, 2000). It indicates whether there is an asymmetry in the distribution of connections. It indicates the degree to which communication is centralized around a single agent or a small group. More centralized groups tend to be more hierarchical in nature.

To calculate the centralization value of an example social network plotted in Figure 1, the following steps are followed:

1.Degree Centrality is calculated.
2.Determine the Maximum Degree of Centrality.
3.Centralization Value: The sum of the differences between the degree centrality of all nodes and the highest degree centrality.

This sum is divided by the theoretical sum of differences that could have the highest degree centrality in the network.

- C_{maks}: The highest degree centrality (centrality of the node with the highest degree).
- C_i: Degree centralities of other nodes.
- The value in the denominator is used as the theoretical maximum centralization value.

The degrees of each node in the network were calculated earlier.

- Node degrees: 3,1,3,4,2,2,3,2,2
- Highest degree: $_{Cmaks}$=4

Calculation of centralization

1.Find the differences between C_{max} - C_i for each node.

Table 4
Calculation of Centralization

c_{maks}- c_i	c_{maks}- c_1	c_{maks}- c_2	c_{maks}- c_3	c_{maks}- c_4	c_{maks}- c_5	c_{maks}- c_6	c_{maks}- c_7	c_{maks}- c_8	c_{maks}- c_9
Difference	4-3	4-1	4-3	4-4	4-2	4-2	4-3	4-2	4-2
Result	1	3	1	0	2	2	1	2	2
Total	14								

2.Calculate the theoretical maximum difference sum. In this case, the theoretical maximum difference occurs when the degree centrality of all nodes except the most central node is zero.

$$maks\sum(C_{maks}-C_i)=(N-1)xC_{maks}=8\times4=32$$

3.Centralization=14/32=0.438

The centralization value of this network is 0.438.

Reciprocity

It is defined as the calculation of whether the connections between nodes are reciprocal, i.e. bidirectional. This metric, calculated in directed networks, is the ratio of the number of node pairs in the network to the number of all possible node pairs. The expression "follow to follow", which is frequently used in social media, refers to the reciprocity metric (Cheng et al., 2011).

SNA Software Tools

Tools for social network analysis are predominantly used for constructing networks, visualizing and manipulating network structures, conducting qualitative and quantitative/statistical analyses, detecting communities, and performing predictive analysis. Despite the availability of many tools, the most widely used ones include Pajek, Gephi, UCInet, NodeXL, R libraries, and NetworkX (Oliveria & Gama, 2011):

- **Pajek**: A free tool designed for analyzing and visualizing large-scale networks.
- **Gephi**: An open-source platform for network manipulation and exploration, featuring a three-dimensional render engine for displaying networks that evolve in real-time.
- **UCInet**: A commercial tool for social network analysis, which uses Pajek and NETDRAW for visualization. It is particularly well-suited for statistical and matrix-based analyses.
- **NodeXL**: A free add-in for Microsoft Excel, providing an accessible, user-friendly way to explore and visualize networks without requiring programming knowledge. However, it is not ideal for analyzing large networks.
- **R libraries** (e.g., igraph, sna, tnet, statnet): Free packages within the R environment, offering a comprehensive set of tools, including a large array of algorithms for community detection, longitudinal network analysis, and two-mode network analysis, with effective two- and three-dimensional visualization options.

NodeXL

NodeXL is mainly used for analyzing networks. It is mostly implemented as an add-in to Microsoft Excel. With the collection of network data, NodeXL provides quick statistics and reporting for people who can use the basic features of Microsoft Excel to analyze network data. NodeXL is a highly effective tool for analyzing and visualizing a social network. In addition to visualizing the entire network in the form of a graph, it can also draw graphs of different social network properties such as Proximity Centrality, Betweenness Centrality, Vertex Degree, Vertex PageRank, etc. Together with Nodexl, it enables network analysis by collecting data from social media platforms such as Twitter, Facebook, YouTube, and Flickr. In addition, topics that are on the agenda on Twitter can be analyzed and analyzed.
The analysis process with NodeXL generally consists of the following steps.

- Importing data
- Data preparation
- Grouping with clustering
- Calculating metrics
- Time series analysis
- Text analysis
- Identifying the Most Important Elements of the Network

- Visualizing the network

GEPHI

GEPHI is an open-source, independent software for visual and network analysis. The primary benefit of utilizing GEPHI for network research is its capability to handle extensive data sets or networks. The GEPHI program possesses certain drawbacks. Occasionally, the response time for a little task or procedure is excessively prolonged. For instance, accessing a file requires considerable time. Given that GEPHI is independent open-source software, it offers numerous functionalities. GEPHI is capable of importing data from text files (TXT), spreadsheets (CSV), and databases. GEPHI is capable of receiving information from various other social network analysis tools. GEPHI facilitates the straightforward graphical representation of a network. GEPHI is capable of producing network graphs and visual representations.

UCINet and Netdraw

UCINet is a menu-based application for social network analysis (Wu & Duan, 2015). UCINet is independent software. In UCINet, all data are represented as matrices. UCINet accepts two categories of input and produces two categories of output. The input comprises input parameters and datasets, whilst the output consists of output text and datasets. The spreadsheet editor is utilized for modifying, inputting new data, and converting UCINet data to Excel or SPSS formats. The UCINet spreadsheet accommodates tiny networks. For extensive datasets, multiple data formats are provided, accessible through an editor known as the dl editor.

UCINet employs Netdraw to facilitate network visualization. Netdraw facilitates various layouts for visualization objectives. These encompass isolates, components, subgroups, and centrality metrics. It also offers functionalities such as node restoration, color scheme adjustment, property visibility toggling, and node size modification, among others.

PAJEK

PAJEK is a software application designed for the visualization and analysis of extensive social networks. It is adequate to calculate most centrality measurements. Furthermore, functions that require many applications can be stored for subsequent re-analysis. PAJEK supports fundamental operations such as subnetwork extraction, identification of linked components (strong, weak, connected), determination of shortest pathways, calculation of maximum flow, centrality assessment (closeness, betweenness, degree, etc.), fragment search, and community detection. The findings generated by PAJEK can be further analyzed with R programming and SPSS. PAJEK accommodates bimodal networks, temporal networks (networks that evolve over time), acyclic and multi-associative networks (many interactions established among the same vertices), and signed networks (networks with both negative and positive connections). PAJEK additionally facilitates text-mining algorithms for the investigation of social networks (Majeed et al, 2020).

NetworkX

NetworkX is a powerful Python library for graph and network analysis. It is used to model and analyze graph structures such as social networks, biological networks, route optimization, connectivity analysis and many more. It allows you to analyze nodes and edges in network or graph structures. It is advantageous for research and applications where new metrics or algorithms need to be developed. Various types of graphs (directed, undirected, weighted, etc.) can be easily calculated with NetworkX. It supports reading and writing data from various formats (e.g. GraphML, GML, Pajek). Together with Matplotlib and other visualization libraries, graph structures can be visualized. In summary, NetworkX is a popular tool for both simple and complex network analysis.

R Programming and Packages

The R programming language has many libraries for analyzing social networks. One

of these libraries is the igraph package. igraph is a fast, efficient, constantly updated, and highly preferred package. Other packages such as sna, tidygraph, and network have a large user base, especially among those interested in statistical network modeling (Gençer, 2023).

Social Network Visualizer

Social Network Visualizer (SocNetV) is a cross-platform, user-friendly free software application for social network analysis and visualization. With SocNetV, the following operations can be performed.

- Draw social networks with a few clicks on a virtual canvas, upload domain data from a file in a supported format (GraphML, GraphViz, Adjacency, EdgeList, GML, Pajek, UCINET, etc.), or browse the internet to create a social network of connected web pages.
- Organize actors and ties through point-and-click, graphs and social network properties that can be analyzed.
- Generate HTML reports.
- Standard graph and network fit metrics such as density, diameter, geodesics and distances, connectivity, eccentricity, clustering coefficient, reciprocity, etc.

VOSviewer and Bibliometrix

The 2 tools used for bibliometric analysis are VOSviwer and Bibliometrix. They are used to process and analyze data extracted from databases such as Scopus or Web of Science. They allow for analysis based on keywords in articles or on the relationships established through authors. Bibliometrix is an R package with a programmable structure (Gencer, 2023).

SNA in Higher Education

Social Network Analysis (SNA) in higher education enables the analysis of relationships and interactions between students, academics, and other stakeholders through network structures. With this method, information flow, collaboration, interaction patterns, and internal dynamics in higher education institutions are analyzed and patterns are revealed. Examples of social network analysis in higher education are listed below.

Examining Student Interactions

Academic achievement and social relationships: How students build social networks inside and outside the classroom can have an impact on their success. The SNA can be used to identify which groups students are more active in and the relationship between academic performance and social interaction.

Improving learning environments: Network structures of interactions in group work, project teams or online platforms (e.g. LMS) are analyzed. It can be revealed which students remain isolated or which groups collaborate more. For example, by analyzing the frequency with which students help each other or share information in the classroom, academic support mechanisms can be better shaped.

Examining Academic and Interdisciplinary Collaboration

Publication and project collaborations: Networks of articles and projects produced by academics together are analyzed. It is revealed which academics play central roles and how interdisciplinary collaborations are shaped.

Strengthening research networks: Areas of intra- or inter-institutional collaboration can be identified and incentives can be given to units that do not have strong ties. For example, by mapping the collaboration network of faculty members in a faculty, joint research projects can be proposed for academics who remain disconnected or isolated.

Analyzing Online Learning Environments

Learning analytics: Students' interactions with each other or with instructors through the LMS (Learning Management System) are analyzed. Student engagement can be increased by analyzing forums, discussion groups, and messaging networks.

Identify students at risk of failure: Isolated students can be identified in advance and counseling and support services can be provided.

Analysis of Management, Leadership, and Organization Networks

Internal decision-making processes: More effective governance can be achieved by analyzing the collaboration and communication structures between administrators, and academic and administrative units.

Leadership and information flow: It can be determined which administrators or academic units are central in information and decision flow, thus optimizing institutional processes.

Alumni tracking systems: Examine how alumni interact through job and career networks. Collaborating with alumni can contribute to the university's career network.

Mentoring networks: By establishing links between alumni and students, students can receive career guidance from alumni.

Policy Development and Performance Measurement

Department and program performance: Strategic plans can be created by examining inter-departmental collaboration networks. In addition, according to the network structures, it is determined which departments need more collaboration.

Innovation and entrepreneurship ecosystems: By analyzing the university's innovation centers and entrepreneurship networks, the ecosystem can be made to work more effectively.

As a result, social network analysis provides a better understanding of the relationships between students, academics, and management structures in higher education. In this way, it is possible to optimize information flow, increase collaboration, and support academic success.

Conclusion and Future Trends

This paper provides a comprehensive overview of the basic principles and objectives of Social Network Analysis (SNA) methods and their applicability to different types of networks. Various SNA metrics and related tasks are described with respect to the different structures of networks. Nowadays, the discovery of network structures in the data generated by many applications and the quality of the information extracted from these networks is increasing the popularity of network analysis. One of the main reasons for this increased interest is the analysis requirements arising from the ever-growing and more complex amount of data. At this point, effectively processing, managing, and extracting meaningful results from large volumes of data flowing at high speed is one of today's most important challenges. Especially with the proliferation of new technologies such as Web 2.0, Internet of Things (IoT), and Industry 4.0, the need for network analysis increases and analysis processes become more demanding and challenging. Therefore, the development of innovative methods that can cope with high volumes of data in network analysis stands out as one of the most current and critical requirements in this field.

Current trends in social network analysis are changing rapidly in line with technological and social developments. Interest and application areas of social network analysis are expanding in areas such as artificial intelligence, data privacy, community analysis, sentiment analysis and language processing, ethical issues, location-based social network analysis, micro-impact analysis, Multilayer Network analysis, disinformation detection, and decentralized networks. These trends indicate that in the future, SNA will become

more effective and have more diverse use cases in both academic and commercial applications.

It is also expected that the popularity of SNA will continue to grow, attracting more researchers to the field and pushing an increasing number of companies to incorporate SNA methods into their business processes and expand their use as strategic tools.

Sample Social Network Analyses

Example 1

Scenario: Let's examine how metrics (degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, local clustering coefficient, density and centralization) are calculated based on students' relationships in a social network of 10 students.

Students: A, B, C, D, E, F, G, H, I, J

Assume that friendship relationships are formed as follows.

- A: B, C, D
- B: A, E
- C: A, D, F
- D: A, C, G
- E: B, H
- F: C, I
- G: D, J
- H: E
- I: F, J
- J: G, I

In the sociomatrix formed according to this network of relationships, each row and column represents an individual. The corresponding cell indicates whether the individuals are connected to each other. It is coded 1 if there is a connection and 0 if there is no connection. The sociomatrix created for individuals whose names are coded as A, B, C, D, E, F, G, H, I, and J is given in Table 5.

Table 5
Relationship Matrix

	A	B	C	D	E	F	G	H	I	J
A	0	1	1	1	0	0	0	0	0	0
B	1	0	0	0	1	0	0	0	0	0
C	1	0	0	1	0	1	0	0	0	0
D	1	0	1	0	0	0	1	0	0	0
E	0	1	0	0	0	0	0	1	0	0
F	0	0	1	0	0	0	0	0	1	0
G	0	0	0	1	0	0	0	0	0	1
H	0	0	0	0	1	0	0	0	0	0
I	0	0	0	0	0	1	0	0	0	1
J	0	0	0	0	0	0	1	0	1	0

The following table shows the formulas for the degrees and the calculated centrality values of each node and the local clustering coefficient.

Table 6
Metric Calculations for Example Network

	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	Clustering Coefficient
Formula	$C_{degree} = \frac{DegreeofNode}{N}$	$c_{betweeness}(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$ $c_{betweeness}(v) = \frac{c_{betweeness}(v)}{(N-1)x(N-2)/2}$	$c_{closeness}(v) = \frac{(N-1)}{\sum_{u \neq v} d(v,u)}$ $c'_{closeness}(v) = \frac{c_{closeness}(v)}{(N-1)}$	$x_i = \frac{1}{\lambda} \sum_{j \in N(i)} x_j$	$c_{cc(i)} = \frac{2x(numberoftriangles)}{dx(d-1)}$
Node					
A	0.33	0.5	0.50	0.49	0.33
B	0.22	0.39	0.41	0.24	0.00
C	0.33	0.28	0.47	0.49	0.33
D	0.33	0.28	0.47	0.49	0.33
E	0.22	0.22	0.32	0.11	0.00
F	0.22	0.17	0.39	0.27	0.00
G	0.22	0.17	0.39	0.27	0.00
H	0.11	0.00	0.25	0.044	0.00
I	0.22	0.06	0.33	0.17	0.00
J	0.22	0.06	0.33	0.17	0.00

The results for the 3 metrics that interpret the whole network (average degree, density and centralization) are as follows.

Average Degree

Average Degree = $\frac{(2xE)}{N}$ = (2x11)/10 =2.2

The average degree of this network is calculated as 2.2. This means that each node in the network has 2.2 links on average.

Density

The density of a network is calculated as the ratio of the number of available edges to the maximum possible number of edges between all nodes in the network.

$$D = \frac{2E}{(Nx(N-1))}$$

E: Number of available edges in the network => E=11

N: Number of nodes in the network => N=10

D= (2*11) / (10*9) = 0.244

The density of this network is calculated to be approximately 0.244. This means that about 24% of the node pairs in the network are directly connected by an edge.

Centralization

The centralization value of the network is a measure of the centrality differences of the nodes in the network and is calculated by the formula.

Centralization = $C = \frac{\sum_{i=1}^{N}(c_{maks} - c_i)}{(N-1)x(N-2)}$

C: Network centralization value

C_{maks}: Degree centrality of the node with the highest degree

C_i: Degree centrality of each node

N: Number of nodes (N=10)

$C_{maks} = 3$

Table 7
(C_{maks} - C_i) Values

	A	B	C	D	E	F	G	H	I	J
C_i	3	2	3	3	2	2	2	1	2	2
C_{maks} - C_i	0	1	0	0	1	1	1	2	1	1

Sum of C_{maks} and C_i differences = 0+1+0+0+1+1+1+2+1+1 = 8

Centralization = C = 8/ (9X8) = 0.11

The centralization value of this network is calculated to be approximately 0.111 or 11.1%. This value indicates that the degree of centralization of the network is quite low and that a central structure between nodes is not very obvious.

Example 2

Let's create a graph of the image for the nodes whose connections are given in Table 8.

Now let us examine the neighborhood matrix (sociomatrix) to show the relationships of the nodes. The number 1 in the cells where the rows and columns intersect indicates that there is a relationship for the intersecting nodes, while the number 0 indicates that there is no relationship.

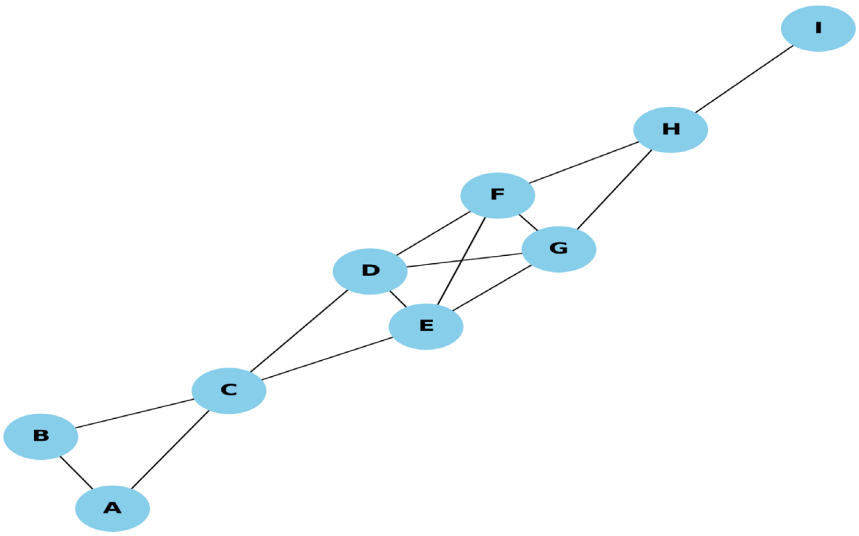
Table 8
Sociomatrix

	A	B	C	D	E	F	G	H	I
A	0	1	1	0	0	0	0	0	0
B	1	0	1	0	0	0	0	0	0
C	1	1	0	1	1	0	0	0	0
D	0	0	1	0	1	1	1	0	0
E	0	0	1	1	0	1	1	0	0
F	0	0	0	1	1	0	1	1	0
G	0	0	0	1	1	1	0	1	0
H	0	0	0	0	0	1	1	0	1
I	0	0	0	0	0	0	0	1	0

For example, when node C is examined, it will be seen that it is related to nodes A, B, D and E.

In order to create this matrix in UCINET program, click on Data>Data Editors>Excel Matrix Editor and in the editor opened by clicking on Data>Data Editors>Excel Matrix Editor, the values seen above should be entered in the rows, columns and cells and saved as Ucinet files (*.##h) as foldername. To create the social network in UCINET program using this matrix, click on Visualize>NetDraw menu and select the previously saved foldername.##h file by clicking the Open button in the NetDraw window. Below is the graph of the network formed according to the connections in the given matrix.

Figure 3
Graph Structure for the Sociomatrix is Given in Table 8



References

Aboelela, S. W., Merrill, J. A., Carley, K. M., & Larson, E. (2007). Social network analysis to evaluate an interdisciplinary research center. *Journal of research administration, 38*(1), 61-75.

Acar, N (2024). Social network analysis in new media research: "Concepts, criteria, approaches". *Journal of Communication Science Research, 4*(2), 167-179. https://doi.org/10.5281/zenodo.10985614

Bloch, F., Jackson, M.O. & Tebaldi, P (2023). Centrality measures in networks. *Social Choice and Welfare, 61*, 413–453. https://doi.org/10.1007/s00355-023-01456-4

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*(5916), 892-895. DOI: 10.1126/science.1165821

Cartwright, D., & Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological review, 63*(5), 277. https://doi.org/10.1037/h0046049

Cheng, J., Romero, D. M., Meeder, B., & Kleinberg, J. (2011, October). *Predicting reciprocity in social networks*. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (pp. 49-56). IEEE.

Codal, K. S., & Coşkun, E. (2016). A network analysis related to the comparison of social network types. *Abant Social Sciences Journal, 16*(1), 143-158. https://doi.org/10.11616/basbed.vi.455794

Coviello, N. E. (2005). Integrating qualitative and quantitative techniques in network analysis. *Qualitative Market Research, 8*(1), 39-60. https://doi.org/10.1108/13522750510575435

DeJordy, R., & Halgin, D. (2008). *Introduction to ego network analysis*. Boston MA: Boston College and the Winston Center for Leadership & Ethics.

Edwards, G. (2010). *Mixed-method approaches to social network analysis.* Discussion Paper. NCRM. Retrived from https://eprints.ncrm.ac.uk/id/eprint/842/

Gençer, M. (2023). *Applied social network analysis*. İzmir Economy University.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory, Vol 1*, 201-233. https://doi.org/10.2307/202051

Laghridat, C., & Essalih, M. (2023). A set of measures of centrality by level for social network analysis. *Procedia computer science, 219*, 751-758. https://doi.org/10.1016/j.procs.2023.01.348

Lewis, J. M. (2011). The future of network governance research: Strength in diversity and synthesis. *Public Administration, 89*(4), 1221-1234. https://doi.org/10.1111/j.1467-9299.2010.01876.x

Majeed, S., Uzair, M., Qamar, U., & Farooq, A. (2020*). Social Network Analysis Visualization Tools: A Comparative Review.* Proceedings-2020 23rd IEEE International Multi-Topic Conference, INMIC 2020.

Marin, A. and Wellman, B. (2011) Social Network Analysis: An Introduction. In Scott, J. and Carrington, P.J., Eds.(pp. 11-25), *The Sage Handbook of Social Network Analysis*. Sage Publications, Thousand Oaks.

Moreno, J. L. (1937). Sociometry in relation to other social sciences. *Sociometry, 1*(1/2), 206-219. https://doi.org/10.2307/2785266

Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(2), 99-115. https://doi.org/10.1002/widm.1048

Scoot, J. (1992). *Social network analysis theory and application*. Newberry Park CA: Sage.

Somyürek, S. & Güyer, T. (2020). Social network analysis. In Güyer, T., Yurdugül, H., & Yıldırım, S. (Eds*), Educational Data Mining and Learning Analytics (*pp.329-375). Nobel.

Streeter, C. L., & Gillespie, D. F. (1992). Social network analysis. *Journal of Social Service Research, 16*(1-2), 201-222. DOI: 10.12691/ajap-1-2-2

Tunalı, V. (2016). *Introduction to social network analysis*. Nobel Publishing.

Tabassum, S., Pereira, F. S., Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(5), e1256. DOI: 10.1002/widm.1256

Wasserman, S. & Faust, K. (2004). *Social network analysis: Methods and applications*. University of Cambridge University Press.

Wu, Y., & Duan, Z. (2015). Social network analysis of international scientific collaboration on psychiatry research. *International Journal of Mental Health Systems, 9*, 1-10.

Zamanitajeddin, N., Jahanifar, M., Bilal, M., Eastwood, M., & Rajpoot, N. (2024). Social network analysis of cell networks improves deep learning for prediction of molecular pathways and key mutations in colorectal cancer. *Medical Image Analysis, 93*, 103071. https://doi.org/10.1016/j.media.2023.103071

About the Author

Akça Okan YÜKSEL received a Ph.D. degree from the Department of Computer Education and Instructional Technologies at Gazi University, one of the most prestigious universities in Türkiye. He works as a Lecturer Doctor in Information System Coordination at Middle East Technical University. His research interests include Social Network Analysis, Learning Analytics, Educational Robotics Applications, Electronic Portfolios, and Learning Management Systems.
E-mail: akca@metu.edu.tr , **ORCID:** 0000-0002-5430-0821

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 17%.

Data Science Applications in Games

Murat ATASOY
Trabzon University

Adil YILDIZ
Trabzon University

Lokman ŞILBIR
Trabzon University

Ekrem BAHÇEKAPILI
Karadeniz Technical University

To Cite This Chapter

Atasoy, M., Yıldız, A., Şılbır, L., & Bahçekapılı, E. (2024). Data Science Applications in Games. In M. Hanefi Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 57-85). ISRES Publishing.

Introduction to Data Science in Gaming

Data science is a multidisciplinary field that uses mathematics, statistics, and computer science techniques to analyze and interpret complex data sets. The primary focus of data science is to extract meaningful insights or meaningful information from data. One of the most important application areas of data science is the digital gaming industry, where large amounts of data are processed and results are customized according to the data. The use of data science techniques in the digital gaming industry is increasing to optimize game design and improve the game environment and player experience. The possibilities provided by data science have become a cornerstone for the industry as game developers, publishers, and marketers (Wallner & Drachen, 2023). The digital game industry uses a vast array of data types to improve both player experience and business outcomes. Some of the data sets used in gaming environments include gameplay data, player behavior data, demographic data, purchasing data, telemetry data, A/B testing data, social data, feedback data, market and competitor data, emotion and biometric data, and VR/AR interaction data.

Game Data

Game data obtained from player actions and interactions in the game environment can provide valuable insights for game developers and analysts (Kriglstein, 2019). For example player movements, clicks, actions, heatmaps, completion rates for levels or quests, in-game purchases and item usage, etc. The collected game data helps developers understand how players interact with the game. This data can be used to fine-tune game mechanics, balance the game difficulty, and improve overall game design.

Player Behavior Data

Player behavior data includes information that includes how players interact with the game environment, patterns, preferences, and habits. Obtaining player behavior directly is a difficult process. Therefore, inferences about player behavior are attempted to be made using a wide variety of data. In this context, data such as session length and frequency, time spent on specific game features or modes, choices made in decision-based games, player retention, and churn rates, as well as social interactions in multiplayer games (such as chatting and team building) are used to determine player behavior. With this player behavior data, game developers can produce data-based solutions for processes such as keeping players in the game, adjusting game difficulty, and creating personalized content for individual users.

Demographic Data

Data obtained regarding the identity of the player using the information provided by players when creating a user account or information collected through surveys is called demographic data. This data includes information such as age, gender, location, game platforms, and device types (Kaye, 2019). This data obtained is important for developers and marketers. Developers and marketers can determine their target audience by analyzing demographic data. They can then make adjustments to the target audience in game features, content, and marketing campaigns. In addition, this demographic data obtained can also be used in the development of localization strategies and platform-specific optimizations.

Purchasing Data

Purchasing data consists of data on virtual currency transactions, product trading patterns, marketplace activity in games with virtual economies, trading, and purchasing behaviors (Cai et al., 2022). Purchasing data helps developers optimize in-game purchasing processes, helping players balance gameplay and purchasing. This helps game developers balance virtual economies and design fair and engaging monetization models.

Telemetry Data

Telemetry data is collected from both the game client and server to monitor performance, player activity, and in-game events (Lim & Harrell, 2014) over distance. This includes many parameters for player behavior, server performance metrics, network latency, connection issues, hardware performance (such as FPS and CPU usage), crash reports, and error logs. Telemetry data is used to optimize game performance, identify technical issues, and ensure smooth gameplay, particularly in online or multiplayer games.

A/B Testing Data

A/B testing data is collected from experiments in which two or more game variants are tested on different player groups. For example, this can involve testing various user interface (UI) designs or button placements, comparing engagement levels for different reward systems or pricing models, and variations of in-game events or levels. A/B testing data enables developers to evaluate which game features or designs perform better with players, leading to more informed decision-making during the development process.

Social Data

Social data is related to player interactions within social features or on external platforms. For example, friends lists, chat data, social invites, team interactions in multiplayer games, social media sharing (such as achievements and screenshots), forums, reviews, and player feedback. Social data assists developers in enhancing multiplayer experiences, community features, and player engagement within the game's ecosystem (Schiller et al., 2019). Additionally, it supports marketing strategies by word-of-mouth or viral spread.

Feedback Data

Feedback data is qualitative data that is gathered from player feedback, reviews, or in-game communication. For example, data gathered from platforms (Steam, App Store, Play Store, etc.), surveys and feedback forms, sentiment analysis from social media or forums, and in-game feedback systems (e.g., post-level ratings). Feedback data makes it easier for developers to understand player satisfaction levels and identify undesirable situations. Positive feedback can enhance intrinsic motivation and long-term play by satisfying competence needs while negative feedback may increase immediate gameplay to improve short-term performance (Burgers et al., 2015).

Emotion and Biometric Data

Emotion and biometric data can be gathered from players' physical responses via emotion recognition software or biometric sensors (Granato et al., 2018). For example, facial expressions and emotional reactions, heart rate, skin conductance or eye movements, and stress or excitement levels can be considered as emotion and biometric data. This data can allow developers to dynamically adjust the game, based on real-time emotional feedback, increasing immersion and personalizing the player experience.

VR/AR Interaction Data

VR/AR Interaction Data is collected from players' interactions with virtual and augmented reality environments. For example hand and head tracking movements, gaze tracking, spatial interaction, and physical gestures type interactions. VR/AR data provides insights into how players interact with 3D environments, helping developers create more immersive and intuitive experiences.

As a result, data collection methods are playing a revolutionary role in the digital gaming industry. Through various data collection methods such as in-game telemetry, A/B testing, player profiling, feedback, and biometric data, developers can optimize game design by better understanding player behavior and preferences.

Data Collection Methods in Gaming

Various methods and approaches are available for collecting game data, including game telemetry, a/b testing, player tracking, event logging, data streaming within games, real-time data collection, game analytics tools, and player profiling techniques. These techniques can have a quantitative or qualitative approach. Game data can focus on development, player satisfaction, game publishing and distribution, and game prediction (Su et al., 2022). Data can be collected live from game-playing sessions and actual players. Examples of in-game data include performance metrics such as in-game movements, interactions, choices, time spent, points levels, and scores, as well as social interactions, economic preferences, bugs and errors, and data related to bots. User comments about games on external platforms are also considered crucial data for developers (D. Lin et al., 2019). These data can provide valuable insights into game satisfaction, enabling the collection and analysis of various aspects related to user contentment.

One of the data collection methods used in game environments is telemetry data. Telemetry data is widely used in game development processes to analyze player behavior, optimize game design, manage monetization strategies, anticipate behaviors, and detect fraudulent activities (Drachen, 2015; Lynn, 2013; Sifa et al., 2018). Using telemetry data, behavioral profiles, and player clusters can be identified. This facilitates the classification of player actions and preferences (Drachen et al., 2014; Kabakov et al., 2014; Sifa et al., 2018). By analyzing telemetry data, developers can predict player engagement by identifying patterns of gameplay and in-game behavior (Bauckhage et al., 2012). On the other hand, by integrating this with traditional user testing protocols, a more comprehensive analysis can be achieved through the effective use of insights into player behavior (Drachen, 2015; Lynn, 2013). Finally, in evaluating the data collected

via telemetry, the application of visualization techniques simplifies the identification of the data, helping to reveal hidden patterns and errors in game design as well as player behaviors (Mirza-Babaei et al., 2014; Moura et al., 2011; Seif El-Nasr et al., 2013).

Developers develop different versions of games, and it may be necessary to conduct statistical studies to reveal the differences between these versions. A/B testing is a data collection and evaluation method used to determine which system version has the best features for specified features (Viljanen et al., 2017). Researchers use A/B testing to gather product/system data and user-centric data ("A/B Testing," 2017; Quin et al., 2023), page load times, and service latency (M. Liu et al., 2019), number of sessions per user, and absence time (Drutsa et al., 2015), playing time, and leveling speed (Drachen et al., 2014) to analyze and draw inferences. A/B tests have hypotheses that claim one situation gets better results. Based on this hypothesis, two game environments/levels/designs are developed and presented to players. With A/B testing, the results obtained from a specific group can be generalized. This can only be made possible with correct sampling. Data science aims to achieve this goal by optimizing factors such as the right sample size, test period, and number of variations (Drovandi et al., 2017). After finding the sample size, advanced data science techniques are used to collect, clean, and process large-scale data (El-Nasr et al., 2021) generated in A/B testing sessions correlated with the hypothesis. Data science provides tools and methods to process these data, provides deep insights, and is useful for measurement and analysis (Gupta & V., 2020). A common method in A/B testing is frequentist inference (Johari et al., 2015), which is based on the assumption that observations are independent and binary or normally distributed. Another method is the Bayesian approach which adds flexibility to studies limited by p significance value. The Bayesian method enables the control of external factors that may affect the results and supports the measurement of differences between variations more effectively (Johari et al., 2015). Probabilistic reward methods (Martín et al., 2021) are used in situations where there is uncertainty and incomplete information on A/B testing. In the context of these and similar data collection and analysis methods, data science contributes to the most effective execution of the process by providing both technical and analytical support in the A/B tests.

In addition to game telemetry and a/b testing, eye tracking and biometric data are also used to collect game data. Eye tracking technology is used to assess emotional responses by analyzing eye movements, fixations, saccades, and pupil diameter during gameplay (J. Z. Lim et al., 2022; Renshaw et al., 2009; Skaramagkas et al., 2023; Tarnowski et al., 2020). Biometric measures such as galvanic skin response (GSR), electrodermal activity (EDA), and heart rate variability are employed to objectively assess emotional activation and engagement levels during gameplay (Ivanina et al., 2023; López-Gil et al., 2016; Simoes & Gomes, 2023; Vazquez et al., 2022). These technologies provide valuable insights into player engagement and emotional activation.

Another data collection method in digital games is event logging. The event log is used to track game errors and performance issues. The data obtained from this process reveals hidden information about player behavior, player performance, and system functionality (Smeddinck et al., 2013; Yu et al., 2022). Analyzing the information in the event logs helps to diagnose errors within the game and enhance the overall gaming experience for players (Cheong & Young, 2006).

As a data collection method, streaming can provide valuable data, particularly regarding the influence of streamers on player engagement. For instance, streaming data from platforms such as Twitch facilitates the capture of player interactions and community dynamics, offering a deeper understanding of player engagement (Micallef et al., 2024). Additionally, data streaming can uncover correlations between streaming activity and in-game performance (Matsui et al., 2020).

Player profiling techniques are an essential tool that encompasses a wide range of methodologies for understanding player behavior and preferences in gaming. For this

purpose, techniques such as behavioral data obtained from in-game logs, surveys and questionnaires, multimodal analysis, and machine learning algorithms are used. One prominent approach involves the utilization of behavioral data obtained from in-game logs, which can reveal patterns in player actions and decision-making processes (Trepte & Reinecke, 2011). Additionally, surveys and questionnaires are frequently employed to gather self-reported data on player motivations, preferences, and psychological states, allowing researchers to correlate these factors with gameplay behavior (Jeong et al., 2020; Trepte & Reinecke, 2011). Moreover, advanced techniques such as multimodal analysis combine behavioral experiments with survey data to evaluate cognitive effects and player experiences (Jeong et al., 2020).

To summarize, various methods for collecting game data include telemetry, player tracking, event logging, and data streaming. Each of these methods provides insights into player behavior and game performance. Telemetry data helps developers to optimize game design and manage monetization strategies, while A/B testing evaluates the effectiveness of different game versions. Eye tracking and biometric measures assess player emotional responses, and event logging tracks performance issues and player interactions. Additionally, data streaming reveals the impact of streaming on player engagement, and player profiling techniques, including surveys and machine learning, help understand player preferences and inform game design. The analysis of gaming data through these diverse methodologies significantly contributes to the understanding of player behavior and the optimization of game design. However, to evaluate the effectiveness of these insights and strategies, game developers and studios must depend on key performance indicators (KPIs) to assess the success of their games.

Game Analytics: Key Metrics and Key Performance Indicators

Key metrics or Key performance indicators (KPIs) are quantifiable metrics to evaluate business, healthcare, and project management-critical initiatives, processes, or objectives. KPIs are instrumental in assessing performance in various sectors, including healthcare to evaluate established standards (Rego et al., 2023), and different industries to form the basis of payments (Lop et al., 2017). One notable application of KPIs is within the gaming industry. In the game industry, key metrics can range from daily and monthly active users (DAU/MAU), retention rate, churn rate, session length, average revenue per user (ARPU), conversion rate, gameplay metrics, virality, player lifetime value, technical performance metrics, engagement metrics, and cognitive and behavioral metrics.

Certain metrics offer insights into player interactions with a game, which are essential for understanding retention rates (Aleem et al., 2016b, 2016a). Another significant engagement metric is session length, which measures the average time players spend in a game during a single session. This metric can indicate how compelling and immersive the game experience is (Koivisto et al., 2023). Also, daily and monthly active users (DAU/MAU) are commonly used to measure the number of unique players engaged with a game over specific time periods and provide insights into the game's popularity and player retention (H. X. Liu & Wagner, 2023).

Monetization strategies are also an important KPI in the gaming industry. In the gaming industry, metrics such as average revenue per user (ARPU), player lifetime value, and conversion rate are commonly used to evaluate the financial success of a game. These metrics help developers understand how much revenue they can expect from each player over time and can also guide decisions regarding pricing models and in-game purchases (Klimas, 2019; Krstić, 2021). The conversion rate is the proportion of free-to-play players who transition into paying users. This metric tracks the percentage of players who make in-game purchases and provides information about the effectiveness of monetization strategies (Junaidi et al., 2018).

Gameplay metrics provide insights into player behavior and performance within the

game. These include metrics such as the number of games played, win/loss ratios, and in-game achievements (Koivisto et al., 2023; Lameman et al., 2010). Satisfaction as an engagement metric can affect games' reputation and longevity. In particular, the fun factor, which evaluates players' overall enjoyment and emotional response, plays a crucial role in understanding this satisfaction (Strubberg et al., 2020). Furthermore, engagement metrics extend beyond individual experiences and include brand loyalty and community engagement, which can be measured through social media interactions, community forums, and player feedback.

Technical performance metrics can be load times, frame rates, memory-CPU-GPU usage, network latency, response time, error rates and crashes, and rendering performance. These metrics provide insights into the game's technical aspects, which significantly influence player satisfaction and engagement. According to Junaidi et al. (2018), load times refer to the duration it takes for a game to start or transition between different levels or scenes, and optimizing these load times is essential for maintaining player engagement. Additionally, network latency, defined as the delay between a player's action and the game's response, is often influenced by internet connection or server speed and can greatly affect gameplay. Junaidi et al. (2018) also emphasize the importance of monitoring errors and crashes during gameplay, as high error rates may indicate underlying issues that need to be addressed to improve the overall player experience. Frame rate indicates how many frames are rendered by the game in one second. A higher frame rate typically results in smoother gameplay, enhancing the player's experience. Conversely, a lower frame rate creates choppy visuals and affects gameplay performance negatively (Koulaxidis & Xinogalos, 2022). Koulaxidis & Xinogalos (2022) also highlight that memory usage tracks the RAM consumed by the game, with high usage potentially causing lag or crashes, especially on resource-limited devices. Similarly, metrics for CPU and GPU (rendering performance) usage indicate the amount of processing, and excessive usage can slow down performance. Tracking both metrics helps developers optimize the game across platforms.

Many metrics, most notably retention, and churn rates are important parameters used in the gaming industry. These indicators also serve as critical inputs for predictive modeling efforts. For example, the ability to predict player churn relies heavily on analyzing in-game behaviors and engagement metrics that KPIs can effectively capture. Being able to predict these values in advance can help game developers make decisions in advance, making games more appealing.

Predictive Modeling in Games

Churn prediction has long been an area of interest in predictive modeling for digital games, especially those using a free-to-play model. In this area, research has been conducted to predict churn with advanced statistical models and machine learning techniques. Roohi et al. (2020) conducted AI-assisted studies to determine the link between game difficulty and player retention. In order to determine player preferences and interests, it is necessary to analyze in-game interactions and player behaviors well. For example, Chen et al. (2018) worked on player activity prediction using time series forecasting with deep learning. This level of application can be used by developers to create game experiences that are customizable and built around the goal of improving player satisfaction and retention. Studies regarding the regularity of player engagement have demonstrated strengthening churn prediction through the mixing of playtime records integration and behavioral data (Yang et al., 2019).

In some studies on churn predictions and player behavior analysis, neural networks, logistic regression and various algorithms were used to analyze. For example, in the game Candy Crush Saga, how each player performed was analyzed with machine learning algorithms, and customizations were made for the game difficulty for each

player. Additionally, Guo's research on player churn prediction emphasized that using different ensemble learning methods showed an increase in prediction performance, as it combined the strengths of multiple models (Guo et al., 2024). These advances in machine learning not only enhance the accuracy of predictions but also provide actionable insights for game developers to improve their strategies.

There are also studies on estimating the outcome in games, that is, the probability of winning. These studies aim to present pre-game and in-game predictions. They are matchmaking systems that focus on matching teams at the beginning of pre-game predictions and try to keep the probability of winning at 50%. These aim to analyze team data, make a win prediction, and match teams with similar values. For example, in DOTA A2, a Multiplayer Online Battle Arena (MOBA) game, telemetry data such as hero selection before the game and post-game end-of-game statistics (score, experience, kill rate) were analyzed with machine learning methods such as Logistic Regression and Random Forest Classifier, and end-of-game predictions were obtained (Kinkade et al., 2015). In another study, data such as time, number of remaining players, team numbers, equipment value, health status, and equipment number from CS-GO in-game data were taken instantly and analyzed with decision trees and Logistic Regression methods, and a dynamic prediction structure was provided (Makarov et al., 2018). Another approach to prediction is the use of real sensor data. In the study conducted by Smerdov et al. (2021), sensor data was processed with the machine learning method and an attempt was made to predict win/lose situations based on player burnout. In the study conducted on the League of Legends game, sensor data for electromyography, eye tracking, seat movements, galvanic skin response, heart rate, facial temperature, electroencephalography and room temperature, humidity, and CO2 levels were used and a structure that can provide feedback for the person to play more defensively in the event of a loss was developed using a model developed with machine learning. Such prediction models improve the gaming experience by helping players understand their situations and the effects of their strategic choices during the game. In the future, it is expected that these models will be applied to more game types and enrich the player experience.

In digital games, predictive modeling is an umbrella term that encompasses a wide variety of high-end statistical and machine-learning tools used to predict player behavior and this research field is very vast. Such integration will consequently allow game developers to enhance player acquisition, improve player retention, optimize game design, and eventually, hit higher revenues. This is a manifestation of the fact that the gaming industry at present stands at an evolutionary crossroads and we can witness likely more increase in the application of predictive models here, hence this area calls for more research efforts and innovation.

Churn, win/lose, matchmaking predictions, and player behavior analysis are deeply interconnected areas within the field of game development, both contributing valuable insights for improving player retention and overall engagement. While prediction focuses on identifying factors that lead players to leave the game or find the fairest match or analyze the user and make suggestions, behavior analysis delves into understanding how players interact with the game environment and detect losing interest, player strategies, key areas, points, or levels.

Player Behavior Analysis

Player behavior analysis in video games utilizes a wide array of methods, providing crucial insights for developers to enhance user experience. These methods can be categorized into several key approaches, including statistical analysis, machine learning, spatial analysis, clustering, and visualization techniques. Each of these methodologies provides unique insights into player behavior, enabling developers to enhance game design, improve player engagement, and foster a more enjoyable gaming experience.

One of the foundational methods for analyzing player behavior is statistical analysis,

which involves the application of various statistical techniques to derive insights from telemetry data. For instance, Bauckhage et al. (2012) emphasized the importance of analyzing distributions of total playing times to understand how players lose interest in games. By examining these distributions, developers can identify critical points in gameplay that may lead to player disengagement, allowing for targeted interventions to retain players. In addition to statistical methods, mixed techniques have also gained prominence in the analysis of player behavior. For example, Canossa et al. (2018) employed a mixed method using telemetry analysis, sequence mining, and clustering to develop detailed player profiles in "Tom Clancy's The Division." using 52 event types and 10.000 players' data. This method allows us to understand usage or gameplay loops, which provide information about player behavior.

By analyzing player behaviors, players can be clustered based on behavioral patterns (Ahmad et al., 2019). These patterns allow for a deeper understanding of player strategies and interactions, which are essential for improving game design and player retention. In the player behavior analysis conducted on 6 million players and 3007 games (Sifa et al., 2014), playing times and game ownership data were examined and cluster analysis showed that only 1/3 of the players distributed their time equally among 3 games, while the rest played a single game. In the same study, it was determined that the playing time was divided into four main clusters, and most games, except for a few, were played for several hours. Additionally, Sifa et al. (2013) applied clustering methods to analyze the behavior of over 62,000 players in "Tomb Raider: Underworld," revealing how player behavior evolves throughout the game.

Building on these techniques, Archetypal Analysis is another method used to examine player data. It is a multivariate data analysis technique that identifies idealized extreme points, or archetypes, to explain observations in a dataset. In the paper by Pirker et al. (2016), which analyzes data from 5000 players and includes 11 activity types such as frequently performed actions, discoveries, points, and time, archetypal analysis is used to understand the evolution of player types and behavior over time and across missions.

Complementing archetypal analysis, spatial analysis offers a deeper insight into player behavior, especially in games with intricate environments. Drachen & Canossa (2011), using data from 28,000 players, introduced spatial analysis of gameplay metrics as a novel approach to user-experience testing. By analyzing the spatial distribution of player actions, developers can identify key areas of interest within the game world. This understanding helps inform both level design and gameplay mechanics, guiding players through the game environment more effectively. In a similar vein, Jim Blackhurst (2011) utilized spatial data from 11.3 million Just Cause 2 players to create heatmaps and 3D visualizations of player death locations, focusing on impact-related deaths (Figure - 1). Overcoming challenges related to large datasets and memory constraints through tools like Processing and OpenGL, Blackhurst was able to render millions of data points in real time. This visualization work enhanced the understanding of player behavior by revealing how players interact with the game environment and identifying significant points of interest.

Figure 1
Visualization of 11.3 Million Players' Death data

Visualization techniques play a crucial role in analyzing player behavior by effectively communicating complex data insights. Moura et al. (2011) observed that traditional methods, such as heatmaps and bar charts, often fall short of representing the temporal progression of player actions. To overcome this limitation, they suggested more dynamic visualization techniques that capture the evolution of player behavior over time. Such methods can greatly enhance developers' understanding of player interactions and inform more strategic design decisions. Additionally, Mirza-Babaei et al. (2014) conducted a case study and proposed a unified visualization approach that integrates qualitative and quantitative data from players' emotional experience from playtesting, offering a more comprehensive understanding of player behavior.

Another significant method for analyzing player behavior is data mining from game telemetry. Lim & Harrell (2014) collected and analyzed 51 types of game metrics including social platform interactions from 219 players to uncover underlying social identity and behavioral patterns. The analysis revealed significant relationships between in-game behavior and social networking interactions, explaining variances in players' number of friends, uploaded screenshots, and videos at rates of 35.1%, 49.6%, and 39.2%, respectively. Furthermore, Hadiji et al. (2014) used twenty million play sessions' telemetry data from five games to predict player churn. The study employs data mining techniques to analyze player behavior and predict churn in freemium games and machine learning algorithms to build predictive models that enhance the understanding of factors influencing player churn.

In conclusion, the analysis of player behavior in video games encompasses a diverse array of methodologies, including statistical analysis, machine learning, spatial analysis, visualization techniques, and user profiling. Each of these methods contributes to a comprehensive understanding of player interactions and preferences, enabling developers to create more engaging and satisfying gaming experiences. By integrating statistical, machine learning, and spatial analysis techniques, developers can not only predict player behavior but also optimize engagement in real-time, ensuring a more cohesive and satisfying gaming experience. To combine these insights, it's clear that integrating adaptive gameplay with advanced analytics offers a powerful toolset to increase player satisfaction. By leveraging methodologies like machine learning and spatial analysis to track player behavior, developers can create dynamic, personalized experiences where players feel a sense of ownership and satisfaction through a tailored gaming experience.

Personalization and Recommendation Systems

In digital games, adaptive gameplay is essential in facilitating personalization and enables the game to change its mechanics and difficulty according to the abilities of

the player. For instance, Zhu & Ontañón (2020) have demonstrated that personalized game experiences can largely grow satisfaction and retention rates among players and so players are more willing to keep playing a game that also adjusts itself to their personal requirements. For example, games like "Minecraft" impel a rather strong sense of ownership and a deep commitment as players are given the option to customize their game experience exactly how they want it.

Adaptive gameplay requires personalization which can give players recommendations of in-game items, tasks, or challenges that suit what they need based on their previous interactions and preferences. This is typically done using collaborative filtering solutions analyzing player data to see what correlations can be made for suggestions. Blocker et al. (2014) discuss how knowing what players like can help with creating better interventions and design strategies, especially for different player types. Through data analytics, developers can craft a more personalized experience that will hook players and keep them playing and investing further. A mobile game, Clash Royale, recommends decks personalized using customized card recommendations based on a player's win-loss history and playstyle preference. In fact, research has shown that games designed with the overall player experience in mind are more likely to provide a degree of control and immersion necessary for long-term engagement (Radhakrishnan et al., 2020). Finally, player feedback can be integrated into the design process serves to improve certain parts of the game and update it accordingly so that it always fits with its audience (Hazar, 2018).

Developers should ensure that these systems keep the privacy and data security of players at paramount importance. A major ethical issue in data collection is transparency around the practice and ensuring that players have a high level of ownership over their data. In closing, personalization and recommendation engines allow digital games to better engage with players by way of adaptive gameplay, dynamic difficult tuning, and tailoring personalized content recommendations. By using collaborative filtering and player-centric design, developers can create games that appeal to specific people, keeping their audience happy and playing longer. As video games continue to evolve, the need for these systems is expected to grow even more in the coming years, meaning constant research and innovation is required to serve this dynamic new player base. As video games increasingly embrace complex customization features, developers and researchers are also focusing on optimizing game mechanics. Leveraging data to balance gameplay elements can increase player satisfaction. Also it supports a fair and engaging experience for all players.

Balancing Game Mechanics Using Data

Data supports game design by helping developers to balance the mechanics of a game. It also guarantees digital game operation fairness, competitiveness as well as player satisfaction. This is a process where data-driven methods are used to change aspects of game play like abilities, items, and level difficulty.

Automatic game balancing can be achieved with deep player behavior models that bring real-time adaptation to player performance and interactions (Pfau et al., 2020). This makes games more fun to play and helps ease the problems with player frustration or disengagement because there are imbalances within the game. Additionally, the way machine learning systems are incorporated into balancing game mechanics could lead to changes around individual player behaviors and preferences.

Another component of balancing game mechanics is matching player skill levels. It is essential in any system that players be matched against others with similar hardware and skill levels to provide fair and competitive matches. More advanced games like Overwatch attempt to match gamers with other players of similar skill levels by analyzing their performance data in-game using a refined matchmaking system. In MOBA games,

pairing inexperienced players with experienced players can frustrate players (Vicencio-Moreira et al., 2015) and lead to an increased churn rate. This is especially critical when it comes to game enjoyment in first-person shooters (FPS) where they often dictate how much fun you can have playing. With the use of predictive algorithms, developers can understand player performance along with skill levels to create a great matchmaking environment that will surely help in retaining players.

Balancing game mechanics is a multifaceted process and requires data-driven balancing, matching player skills, and extensive balance testing methodologies. These tactics allow a lot of game devs to build better games that are more fun for many different players at higher and lower levels. As the gaming industry keeps changing, it will become increasingly important to think with data when balancing games, and there is a need for more research and innovation in this field.

Fraud Detection and Anti-Cheating Measures

Integrity in online gaming environments absolutely depends on protecting digital games from fraud and from cheating. With the rise of digital games, though, also comes a rise in cheating and other malicious actions that can shake the player's faith and deteriorate gameplay as a whole. One of the keys to maintaining a safe and clean environment in any type of digital game is enabling game security analytics to detect, find out, and suppress cheating. Analyzing player data and observing abnormalities in game play that may reveal cheaters, for example, the online first-person shooter game Counter-Strike: Global Offensive (CS-GO) uses complex behavior analysis to monitor player movements and actions for signs of cheating. Chapel et al. (2010) focus on detection techniques for cheating and talk about the need to base these in robust data-driven algorithms to identify suspect behavior and take appropriate action, using probabilistic methods as well. Such analyses can be used to provide a picture of how frequent and what kinds of cheating there is, which in turn helps developers direct their resources towards making the game more secure.

In digital gaming fair play is relatively important for which anti-cheat systems are indispensable. Utilizing a variety of methods to identify and mitigate cheating, these systems monitor player behaviors, analyze in-game data logs and implement machine learning algorithms. Alayed et al. (2013) presents a set of machine learning based cheat detection algorithms to recognize cheating patterns in internet First-Person Shooter games according to the player behaviors. These systems not only catch known cheats, but they continuously learn new ways to cheat so that the integrity of the system is maintained. Anomaly detection is another important part of cheating in digital games. By defining player behaviors over a reasonable span of time, developers can identify abnormalities suggestive of cheating. Sophisticated algorithms in games like Call of Duty: Warzone will monitor where players are moving and what weapons they have equipped, which can then flag oddities as potential cheats. This emphasis is especially valuable in multiplayer structures where one player's actions can greatly affect the experiences of others. S. J. Lee et al. (2021) argue that knowledge of these psychological issues could be made use of to enhance anomaly discovery initiatives by highlighting the impact of competitive motivation and self-training concerning what propels individuals to rip off habits. The experience is that developers can learn more about the psychology of cheaters and become better data analysts.

One key aspect of their anti-cheat solutions is bot detection, as one of the most common forms of cheating is automated scripts or bots that take over character gaming actions resulting in broken gameplay and unlevel playing field for normal non-pro gamers. McDaniel & Yampolskiy (2012) provide an example of such security for online games, the embedded CAPTCHA element prevents bot activity thus ensuring game integrity through technical solutions. The researchers also outline the use of machine learning in

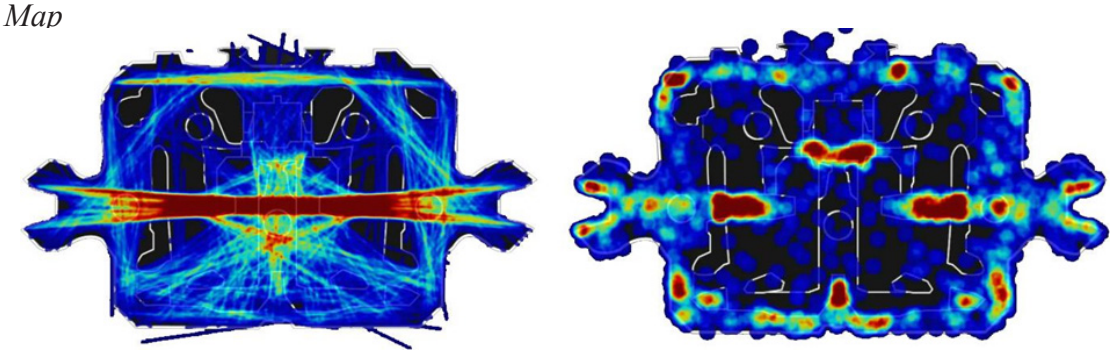
identifying first-person shooter bots, highlighting how these technologies can provide a sophisticated and effective approach to stopping cheating (Kanervisto et al., 2023). In summary, cheat detection and anti-cheat mechanisms in digital games are a complex multi-level process involving digital game security analytics, coupled with cheat-hacks/bots detection, anomaly identification, and bot networks tracking along with enforcement of fair play. Using data-driven methods and tools available in libraries can help developers in tackling cheating and ensuring that games stay fun. With the gaming ecosystem changing and maturing, further study and innovation in these fields will be necessary to combat upcoming fraud and cheating threats.

Data Visualization for Game Development

The visual system of humans has the most dedicated cells in the brain, reflecting its extensive neural resources allocated to processing visual information (Nassi & Callaway, 2009). This shows that people can perceive visual data faster than data from other sensory sources. At this stage, data visualization is a tool that makes it easier for people to perceive complex data (Silva, 2016; L. Zhou, 2023). Through data visualization, complex data transformed into visual formats can be easily perceived thanks to the innate abilities of the human visual system (Healey & Enns, 2012). Data Visualization is a subfield of visualization that focuses on the graphical representation of data. By transforming raw data into visual formats such as charts, graphs, and maps, data visualization makes it easier for people to see patterns, trends, and outliers (Gerela et al., 2022). Such as Figure 2, created by Halo® Game Data to visualize the sightline and kill of beam rifles on the Coliseum map (Halo Heatmaps, 2016). With this visualization, users can understand where to use or avoid the beam rifle and change their in-game strategies accordingly. This makes it easier to better analyze and communicate the information. Data visualization is an important component of data science that enables the discovery, analysis, and communication of complex data (Govind Shinde & Shivthare, 2024; Keim et al., 2013). Data visualization is one of the indispensable tools for a data scientist because it makes raw data easily observable.

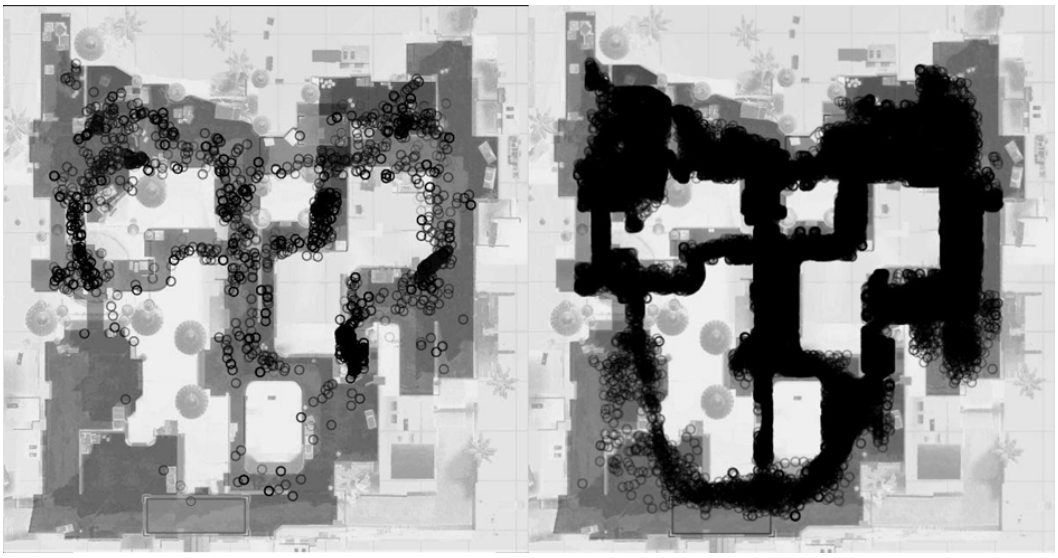
Figure 2

Visualization of the sightline (left) and death (rigth) map of a gun on Halo - Coliseum Map



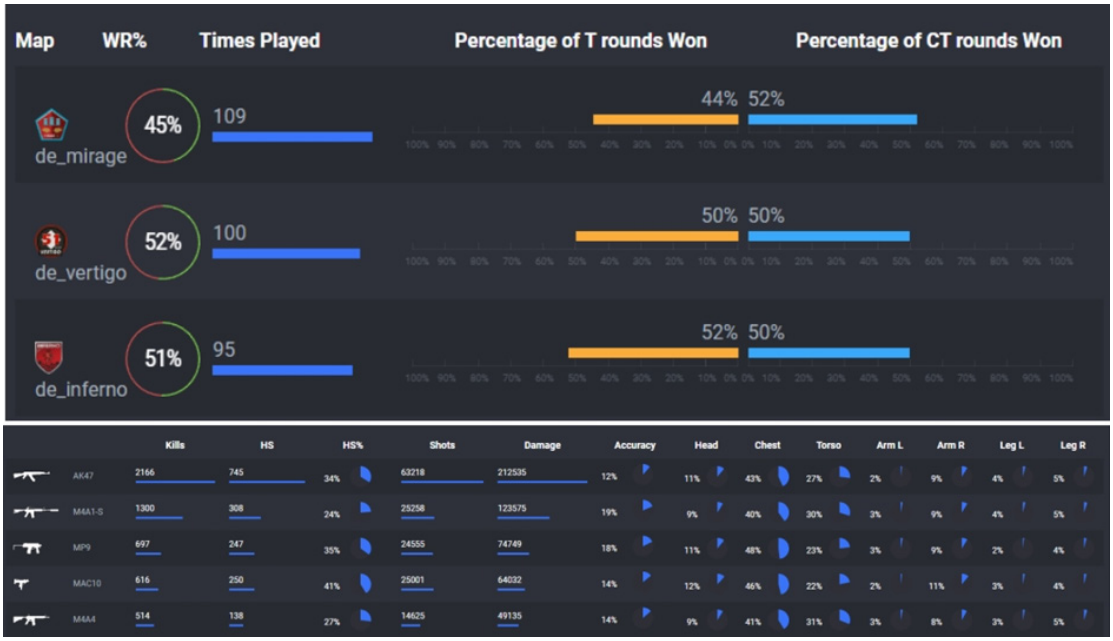
It is also possible to frequently come across data visualization applications in games. Data visualization can be considered in two ways in games. The first of these is the visualization elements that are directly included in the game, such as providing feedback to the user about the game and information about the current game situation. The other is the visualization applications aimed at providing game developers with insight into the user's gaming experience. In Figure 3, data visualization shows the distribution of thrown grenades on a map (Jacobwdym, 2018) over more than 410,000 played rounds in the competitive game Counter-Strike: Global Offensive (CS-GO).

Figure 3
1000 samples of thrown grenades on the left, whole data on the right (Jacobwdym, 2018)



Data visualizations in the game provide the player with a quick and easy-to-understand summary of their status through methods such as dashboards, information panels, and leaderboards, helping them evaluate their current performance and make strategic decisions (Meyer & Bishop, 2022; Ruiperez-Valiente et al., 2021). Thanks to these visualizations, players can monitor their success in the game, notice their areas of development, and plan their future moves more consciously (Danak & Mannor, 2011). In Figure 4, a user's CS-GO gameplay data is displayed, allowing the user to evaluate their performance on three maps and decide which one to improve. Additionally, the user can determine which weapon is best for their gameplay by reviewing their statistics.

Figure 4
CS-GO user game statistics with visuals (CSSTATS (ESL Gaming Online), 2024).



For game developers, data visualization enables the development of better games, in other words, user-friendly designs, by enabling the analysis of how the user behaves and interacts in the game, where they focus more, and the order of the steps in the process (Kriglstein, 2019). Similarly, in order to develop gamification, which is one of the most

critical elements in games, data visualization is used to analyze behavioral data in games by combining elements such as points, levels, leaderboards, badges, and challenges to increase user participation and contribution (Yampray & Inchamnan, 2019). Especially visualization of event sequences in games facilitates the understanding of players' skill development, action and timing patterns, and strategy changes, allowing the design of game stages and the examination of player groups (Li et al., 2019). Finally, while data visualization plays an important role in enriching gaming experiences, additional machine learning and AI analytics may need to be implemented to minimize data privacy and misinterpretation risks. With the support of these technologies, more accurate and reliable analytics can be provided, enabling responsible and balanced game design.

Machine Learning Applications in Gaming

In the context of machine learning applications in gaming, advances in areas such as AI-powered NPCs, procedural content generation, game AI, reinforcement learning, and deep learning are enabling gaming environments to become more dynamic, adaptive, and personalized. These technologies not only enhance the complexity of in-game mechanics but also offer novel approaches to content creation and player interaction.

One of the application areas of machine learning in games is creating AI-supported Non-Player Characters (NPCs). Data science plays a pivotal role in the development of NPCs that can adapt to players in real time, enhancing the gaming experience through intelligent behavior and dynamic interactions. The integration of machine learning techniques, particularly reinforcement learning where an agent learns to make decisions by interacting with an environment to maximize cumulative rewards (Merrick, 2008), has been instrumental in creating NPCs that exhibit believable and adaptive behaviors. For example, Cruz and Uresti state that well-developed NPCs that react indistinguishably from human players are crucial to an immersive gaming experience (Cruz & Uresti, 2018). In addition, with the ML-Agent module on Unity (Juliani et al., 2018), and the Learning Agents module on Unreal Engine (Brendan Mulcahy & Daniel Holden, 2023), NPCs can be humanized and made intelligent in the desired direction. This adaptability is further supported by the use of behavior trees, which provide a flexible framework for modeling NPC logic, allowing for more natural interactions and decision-making processes (Kozik et al., 2021). There are two main methods for developing AI-based NPCs, namely task-oriented learning and game-oriented learning (Kaur et al., 2023). Task-driven learning involves agents learning to complete specific tasks in controlled environments based on user-defined goals and action sequences. Game-driven learning occurs in game-like settings where agents learn strategies to win or survive, following the game's rules and constraints.

Another Machine Learning Application used in games is Large language models (LLMs). LLMs can be used to provide real-time feedback and adaptation of game mechanics based on player behavior and avatar interactions. This adaptability can lead to a more engaging gaming environment, where players feel that their actions and choices have significant consequences, further deepening their connection to their avatars and the game world (Taesiri et al., 2022). LLMs can enable NPCs to generate real-time meaningful dialogue, automate narrative creation (Kumaran et al., 2023), personalized gaming experiences, exhibit emotional responses (Garavaglia et al., 2022) and personalities for nuanced interactions (Karpouzis & Tsatiris, 2022), learn and evolve from player interactions (Giunchi et al., 2024), and proactively engage players (Giunchi et al., 2024; Sun et al., 2023) by assigning quests or providing contextual information.

Procedural content generation (PCG) enhances the game development process by making it more dynamic (Bernardi et al., 2021), decreases development time, and makes the in-game experience more adaptive and more replayable (Van Linden et al., 2013). Levels (Z. Zhou & Guzdial, 2021), quests (Soares et al., 2016), and even music (Plans & Morelli, 2012) can be created dynamically through PCG. PCG also helps in reducing the

size of game files on the hard drive for customers (Summerville et al., 2017). In his study, Summerville created the machine learning data prepared for PCG through the analysis of the cards that could be collected, level data obtained by frame-by-frame analysis of gameplay videos, and various fictional stories. Level data is seen to be a frequently used data type for machine learning-supported PCG (Khalifa et al., 2020). While 2D game levels are dominant, 3D game levels, story text, rhythm, character models, textures, and cards are also produced by PCG.

In order for games to become more enjoyable and reduce the churn rate, the matchmaking of users must be balanced and competitive. AI can analyze player behavior to match players with similar playstyles. To manage this, the player's historical data can be analyzed to understand the player's behavior, and using this data user levels-styles can be determined (Sapienza et al., 2017). Most of the applications that aim to balance game mechanics and game prediction in games reveal the use of artificial intelligence with machine learning and reinforcement learning methods. Together, machine learning and big data are transforming gaming by improving personalization and dynamic in-game experiences. Machine learning models, like those powering NPC behavior and procedural content generation, generate vast amounts of player data that big data analytics then processes to improve engagement and enhance game mechanics. This combination allows developers to make data-driven adjustments and optimize gameplay based on real-time insights into player preferences and trends.

Big Data Challenges in the Gaming Industry

The concept of big data has gained popularity after the widespread use of the internet and mobile devices, which are vital for daily life. Social media platforms, internet applications, and sensors that obtain data directly from individuals enable the production of large amounts of data and the transformation of the produced data into products. The comprehensive, diverse, and constantly growing data sets contained in big data are quite difficult to process and manage using traditional methods (I. Lee, 2017; Munawar et al., 2020). The main purpose of big data is to obtain general insights from comprehensive data sets and use this information in processes such as data-driven decision-making, predictive analysis, personalized experiences, productivity, and innovation (Sahoo, 2022; Wu et al., 2014).

The increase in the use of the internet and mobile devices has also rapidly expanded the gaming industry, which has a large follower base. The rapid growth of the game sector has also led to an increase in the data produced in games. The main reasons for this are the combination of factors such as the increasing online games, in-game interactions, advanced graphics, the increase in multiplayer games due to the improvement of the internet infrastructure, the increase in the game-playing capacity of mobile devices, and the use of artificial intelligence (Bauckhage et al., 2015; García-Álvarez et al., 2017; Wallner & Drachen, 2023).

Managing and using this big data generated in games has various challenges, primarily due to the enormous volume, variety, and speed of the data generated. For example, companies like Zynga process billions of rows of data every day, which complicates the data management process and analysis (Reynolds, 2019). The diversity of structured and unstructured data generated complicates the extraction of meaningful results, as traditional data management systems have difficulties in accommodating such heterogeneity (Cui et al., 2020). However, as the amount of data generated due to the growing use of online games increases, it is necessary to develop strong infrastructure and analysis processes to ensure timely decision-making (C. H. Lee & Yoon, 2017). At this point, the inherently complex structure of big data can lead to data quality issues such as completeness and accuracy, which are critical for effective analysis (Bandara et al., 2024).

In conclusion, big data technologies significantly enhance the gaming industry by

enabling data-driven decision-making and improving user engagement. For example, big companies with millions of users use extensive analytics to understand player behavior and optimize gaming experiences by processing billions of rows of data daily (Reynolds, 2019). Thanks to this data-centric approach, the results developers obtain from game analytics allow them to effectively adapt game features and marketing strategies (Mäntymäki et al., 2020; Su et al., 2022). Additionally, small and medium-sized game developers leverage analytics to improve game designs and revenue forecasts, emphasizing the democratization of data access in the industry (Mäntymäki et al., 2020; Su et al., 2022). On the other hand, the integration of big data encourages innovation by allowing developers to predict trends and user preferences, thereby improving overall gaming sustainability (Na et al., 2022). The rapid increase in data production in the gaming sector has presented both opportunities and challenges. Big data technologies enable developers to gain insights into player behavior, optimize the gaming experience, and drive innovation. However, this also raises ethical and privacy concerns, as extensive data collection and analysis can lead to biased results and privacy risks.

Ethics and Privacy Concerns in Game Data Science

Game data science leverages the amount of data generated by digital games to improve various aspects of the game development process (Seif El-Nasr, 2019). As in any field that processes large amounts of user data, ethical considerations are very important in this field. One of the primary ethical concerns in-game data science is the potential for bias in data-driven decision-making. The datasets used to train algorithms and build models in the game industry may not accurately reflect the full spectrum of player demographics, behaviors, and preferences, potentially leading to biased outcomes that unfairly impact certain groups (Kuhlman et al., 2020). There is a risk of biased data, where certain populations are underrepresented or certain attributes like race, gender, and age are not evenly distributed (Schneider et al., 2023). Therefore, if the developed algorithms are trained on biased data sets, they may unintentionally lead to applications with features aimed at certain groups (Seif El-Nasr & Kleinman, 2020). To address these issues, game developers and data scientists must take proactive steps to ensure that their data-driven practices are transparent and accountable. Players are often unaware of the extent to which their data is collected and used. Based on the information gathered, the ethical considerations in-game data science are not merely an extension of general data ethics but require specific attention due to the unique nature of gaming environments. To solve these problems, consideration should be given to the use of digital approval processes (Charles & Magtanong, 2021).

Game developers and data scientists must work together to create ethical frameworks that consider player behavior and the potential impact of data-driven decisions on the gaming experience. Furthermore, it is essential to foster a culture of ethical awareness within the gaming industry, ensuring that all stakeholders understand the importance of ethical practices in-game data science. To address these concerns, researchers propose developing ethical guidelines, and increasing transparency and accountability (Seif El-Nasr & Kleinman, 2020; Melhart et al., 2024).

Companies collect and use player data to generate revenue and provide better experiences. This situation brings privacy concerns in games (Laakkonen et al., 2016). Modern gaming platforms gather extensive personal information through various means, including hardware sensors, social features, and tracking technologies (Russell, Reidenberg & Moon, 2018). This data collection process raises ethical and privacy concerns about user security, particularly for child gamers (Dasgupta & Sarkar, 2022). Studies show that transparency regarding data practices in the gaming industry is often lacking, especially concerning third-party sharing (Russell, Reidenberg & Moon, 2018). Gamers consider the biggest threats to their privacy while gaming to be the exposure of sensitive information, including passwords, location data, and purchasing or financial

details. This data can be collected in-game environments without the knowledge of users, especially through microtransactions (Dasgupta & Sarkar, 2022). To address these issues, researchers suggest incorporating privacy-sensitive approaches into game platform design (Laakkonen et al., 2016) and improving user control mechanisms and privacy settings (Russell, Reidenberg & Moon, 2018).

On the other hand, the integration of AI in gaming brings new ethical challenges, necessitating responsible AI practices (Canca et al., 2024). The gaming industry's growing influence and access to resources come with social responsibilities that have often been neglected (Cook, 2021). Ethical concerns in AI-driven games include the artificial induction of emotions, privacy issues in creating safe gaming spaces, and challenges to transparency and ownership in the game environment (Melhart et al., 2024). To address these challenges, the gaming industry needs to adopt responsible AI practices, tools, and governance structures (Canca et al., 2024). As the ethical and privacy aspects of game data science continue to grow in importance, the integration of AI and advanced data processing methods is making the process more complex and providing new opportunities. At the same time, emerging trends such as edge computing, "living games" with ever-evolving NPCs, and blockchain-based assets are creating a dynamic space that enhances user experience and data-driven interaction. These innovations and more will shape the next generation of immersive, personalized, and emotionally intelligent gaming experiences, where ethical frameworks, AI applications, and innovative technologies converge.

Future Trends: AI and Data Science in Gaming

Edge computing offers solutions to latency and bandwidth problems in gaming environments by making computing resources accessible to users (Hammad et al., 2023). Real-time Edge computing is particularly useful for mobile augmented reality (AR) games that require high responsiveness and processing power (Hammad et al., 2023). Besides, edge computing is also useful for video streaming that requires high bandwidth and low latency (Bilal & Erbad, 2017). Applications such as real-time video analytics, vehicle applications and multi-user cloud games, and smart city technologies will benefit more from edge computing in the future (L. Lin et al., 2019). These developments can improve user experience and open up new possibilities in gaming environments.

Another concept that data analysis and AI have revealed in-game environments is the concept of "living games". The concept of "living games" with dynamic Non-Playable Characters (NPCs) that evolve, and interact in real time, is becoming increasingly feasible thanks to advancements in AI and procedural content generation (Cruz & Uresti, 2018). Generative AI models like GPT-4 enable more immersive experiences by allowing NPCs to engage in real-time, unscripted conversations. For instance, systems like PANGeA use generative AI to allow NPCs to develop personalities and exhibit human-like traits based on psychological models. This enables NPCs to continue interacting and evolving even when the player is not in the game, making the game world feel more alive. Players might return to the game to find that NPCs have developed new relationships or knowledge based on their own activities, independent of player intervention (Buongiorno et al., 2024).

Another technology that is expected to increase its impact in the gaming world is blockchain technology. Thanks to blockchain, a more secure and transparent process can be provided in the management of in-game assets of NFTs (Non-Fungible Tokens). NFT-based in-game assets can provide users with a greater sense of ownership of digital assets and therefore digital games by ensuring that each asset is unique and its ownership can be verified on the blockchain. (Paajala et al., 2022; Paduraru et al., 2022). Thanks to the infrastructure provided by blockchain, it is also easier to create a shared marketplace and digital wallet. Such marketplaces can allow players to securely buy and sell the digital

assets they earn or obtain, and even move them between different games (Paduraru et al., 2022). Play-to-earn models, in particular, allow players to directly convert in-game assets and services into money. In this way, the gaming experience can be evaluated not only for entertainment but also as economic gain. These opportunities offered by Blockchain will also provide game developers with the opportunity to create new business models. Developers can leave control of the assets in the game to the players, allowing the assets to be traded even in markets outside the game. Such developments will enable the expansion of game ecosystems to a wider universe and improve player experiences (Paajala et al., 2022).

On the other hand, studies are being conducted to investigate the nature of the emotions evoked by video games, how emotions are produced through games, and how games can be used for emotion regulation (Hemenover & Bowman, 2018). Studies have suggested conceptual frameworks for emotional design and emphasized the importance of understanding emotions in order to create interesting environments (de Byl, 2015; Hemenover & Bowman, 2018). Challenges in this field include developing intelligent systems capable of accurately interpreting players' emotional states and adjusting game narratives accordingly (Kotsia et al., 2013). Looking forward, Kotsia et al. (2013) argue that affective gaming holds the potential to transform player-game interactions. Imagine a game that can sense when a player is feeling frustrated and automatically offers assistance, or one that increases the emotional intensity of a scene based on the player's response. Moreover, as more studies focus on the relationship between games and emotions, the development of emotionally intelligent games could lead to innovations in game design, creating more empathetic and emotionally resonant gaming worlds. Thus, the future of emotional engagement in games will not only enhance entertainment but also support emotional growth, offering new ways for players to connect with games on a personal and emotional level.

References

A/B Testing. (2017). In *Encyclopedia of Machine Learning and Data Mining*. https://doi.org/10.1007/978-1-4899-7687-1_100507

Ahmad, S., Bryant, A., Kleinman, E., Teng, Z., Nguyen, T. H. D., & El-Nasr, M. S. (2019). Modeling Individual and Team Behavior through Spatio-temporal Analysis. *CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. https://doi.org/10.1145/3311350.3347188

Alayed, H., Frangoudes, F., & Neuman, C. (2013). Behavioral-based cheating detection in online first person shooters using machine learning techniques. *IEEE Conference on Computatonal Intelligence and Games, CIG*. https://doi.org/10.1109/CIG.2013.6633617

Aleem, S., Capretz, L. F., & Ahmed, F. (2016a). A Digital Game Maturity Model (DGMM). *Entertainment Computing*, *17*. https://doi.org/10.1016/j.entcom.2016.08.004

Aleem, S., Capretz, L. F., & Ahmed, F. (2016b). Empirical investigation of key business factors for digital game performance. *Entertainment Computing*, *13*. https://doi.org/10.1016/j.entcom.2015.09.001

Bandara, F., Jayawickrama, U., Subasinghage, M., Olan, F., Alamoudi, H., & Alharthi, M. (2024). Enhancing ERP Responsiveness Through Big Data Technologies: An Empirical Investigation. *Information Systems Frontiers*, *26*(1). https://doi.org/10.1007/s10796-023-10374-w

Bauckhage, C., Drachen, A., & Sifa, R. (2015). Clustering Game Behavior Data. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(3). https://doi.org/10.1109/TCIAIG.2014.2376982

Bauckhage, C., Kersting, K., Sifa, R., Thurau, C., Drachen, A., & Canossa, A. (2012).

How players lose interest in playing a game: An empirical study based on distributions of total playing times. *2012 IEEE Conference on Computational Intelligence and Games, CIG 2012*. https://doi.org/10.1109/CIG.2012.6374148

Bernardi, A., Gadia, D., Maggiorini, D., Palazzi, C. E., & Ripamonti, L. A. (2021). Procedural generation of materials for real-time rendering. *Multimedia Tools and Applications*, *80*(9). https://doi.org/10.1007/s11042-020-09141-9

Bilal, K., & Erbad, A. (2017). Edge computing for interactive media and video streaming. *2017 2nd International Conference on Fog and Mobile Edge Computing, FMEC 2017*. https://doi.org/10.1109/FMEC.2017.7946410

Blocker, K. A., Wright, T. J., & Boot, W. R. (2014). Gaming preferences of aging generations. *Gerontechnology*, *12*(3). https://doi.org/10.4017/gt.2014.12.3.008.00

Brendan Mulcahy, & Daniel Holden. (2023, March 31). *Learning agents introduction*. https://dev.epicgames.com/community/learning/tutorials/8OWY/unreal-engine-learning-agents-introduction

Buongiorno, S., Klinkert, L. J., Chawla, T., Zhuang, Z., & Clark, C. (2024). PANGeA: Procedural Artificial Narrative using Generative AI for Turn-Based Video Games. *ArXiv Preprint*.

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*, 94–103. https://doi.org/10.1016/j.chb.2015.01.038

Cai, X., Cebollada, J., & Cortiñas, M. (2022). A grounded theory approach to understanding in-game goods purchase. *PLOS ONE*, *17*(1), e0262998. https://doi.org/10.1371/journal.pone.0262998

Canca, C., Ihle, L. H., & Schoene, A. M. (2024). Why the Gaming Industry Needs Responsible AI. *ACM Games*, *2*(2). https://doi.org/10.1145/3675803

Canossa, A., Makarovych, S., Togelius, J., & Drachen, A. (2018). Like a DNA string: Sequence-based player profiling in Tom Clancy's the Division. *Proceedings of the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2018*. https://doi.org/10.1609/aiide.v14i1.13049

Chapel, L., Botvich, D., & Malone, D. (2010). Probabilistic approaches to cheating detection in online games. *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, CIG2010*. https://doi.org/10.1109/ITW.2010.5593353

Charles, W., & Magtanong, R. (2021). *Ethical Benefits and Drawbacks of Digitally Informed Consent*. https://doi.org/10.4018/978-1-7998-8467-5.ch008

Chen, P. P., Guitart, A., Del Río, A. F., & Periáñez, A. (2018). Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*. https://doi.org/10.1109/BigData.2018.8622151

Cheong, Y. G., & Young, R. M. (2006). A framework for summarizing game experiences as narratives. *Proceedings of the 2nd Artificial Intelligence and Interactive Digital Entertainment Conference, AIIDE 2006*. https://doi.org/10.1609/aiide.v2i1.18754

Cook, M. (2021). The Social Responsibility of Game AI. *IEEE Conference on Computatonal Intelligence and Games, CIG*, *2021-August*. https://doi.org/10.1109/CoG52621.2021.9619090

Cruz, C. A., & Uresti, J. A. R. (2018). HRLB^2: A reinforcement learning based framework for believable bots. *Applied Sciences (Switzerland)*, *8*(12). https://

doi.org/10.3390/app8122453

CSSTATS (ESL Gaming Online). (2024). *Player statistics | CS2 Stats*. https://csstats.gg

Cui, Y., Kara, S., & Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. In *Robotics and Computer-Integrated Manufacturing* (Vol. 62). https://doi.org/10.1016/j.rcim.2019.101861

Danak, A., & Mannor, S. (2011). A robust learning approach to repeated auctions with monitoring and entry fees. *IEEE Transactions on Computational Intelligence and AI in Games*, *3*(4). https://doi.org/10.1109/TCIAIG.2011.2160994

Dasgupta, Dr. D., & Sarkar, S. (2022). Privacy: A myth in online gaming? *International Journal of Advanced Mass Communication and Journalism*, *3*(2). https://doi.org/10.22271/27084450.2022.v3.i2a.49

de Byl, P. (2015). A conceptual affective design framework for the use of emotions in computer game design. *Cyberpsychology*, *9*(3). https://doi.org/10.5817/CP2015-3-4

Drachen, A. (2015). *Behavioral Telemetry in Games User Research*. https://doi.org/10.1007/978-3-319-15985-0_7

Drachen, A., & Canossa, A. (2011). Evaluating motion: Spatial user behaviour in virtual environments. In *International Journal of Arts and Technology* (Vol. 4, Issue 3). https://doi.org/10.1504/IJART.2011.041483

Drachen, A., Thurau, C., Sifa, R., & Bauckhage, C. (2014). *A Comparison of Methods for Player Clustering via Behavioral Telemetry*.

Drovandi, C. C., Holmes, C. C., McGree, J. M., Mengersen, K., Richardson, S., & Ryan, E. G. (2017). Principles of experimental design for Big Data analysis. *Statistical Science*, *32*(3). https://doi.org/10.1214/16-STS604

Drutsa, A., Gusev, G., & Serdyukov, P. (2015). Future User engagement prediction and its application to improve the sensitivity of online experiments. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*. https://doi.org/10.1145/2736277.2741116

El-Nasr, M. S., Dinh, T. H. N., Canossa, A., & Drachen, A. (2021). Game Data Science : An Introduction . In *Game Data Science*. https://doi.org/10.1093/oso/9780192897879.003.0001

Garavaglia, F., Nobre, R. A., Ripamonti, L. A., Maggiorini, D., & Gadia, D. (2022). Moody5: Personality-biased agents to enhance interactive storytelling in video games. *IEEE Conference on Computatonal Intelligence and Games, CIG*, *2022-August*. https://doi.org/10.1109/CoG51982.2022.9893689

García-Álvarez, E., López-Sintas, J., & Samper-Martínez, A. (2017). The Social Network Gamer's Experience of Play: A Netnography of Restaurant City on Facebook. *Games and Culture*, *12*(7–8). https://doi.org/10.1177/1555412015595924

Gerela, P., Mishra, P. N., & Vipat, R. (2022). Study on data visualization. *International Journal of Health Sciences*. https://doi.org/10.53730/ijhs.v6ns3.7393

Giunchi, D., Numan, N., Gatti, E., & Steed, A. (2024). DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming. *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 579–589. https://doi.org/10.1109/VR58804.2024.00078

Govind Shinde, Ms. B., & Shivthare, Dr. S. (2024). Impact Of Data Visualization In Data Analysis To Improve The Efficiency Of Machine Learning Models. *Journal of Advanced Zoology*. https://doi.org/10.53555/jaz.v45is4.4161

Granato, M., Gadia, D., Maggiorini, D., & Ripamonti, L. A. (2018). Feature Extraction

and Selection for Real-Time Emotion Recognition in Video Games Players. *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 717–724. https://doi.org/10.1109/SITIS.2018.00115

Guo, R., Xiong, W., Zhang, Y., & Hu, Y. (2024). Enhancing Game Customer Churn Prediction with a Stacked Ensemble Learning Model. *The Journal of Supercomputing*, *Preprint*. https://doi.org/10.21203/rs.3.rs-4333328/v1

Gupta, P., & V., N. (2020). The Impact and Importance of Statistics in Data Science. *International Journal of Computer Applications*, *176*(24). https://doi.org/10.5120/ijca2020920215

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., & Bauckhage, C. (2014). Predicting player churn in the wild. *IEEE Conference on Computatonal Intelligence and Games, CIG*. https://doi.org/10.1109/CIG.2014.6932876

Halo Heatmaps. (2016). *Halo Heatmaps*. https://haloheatmaps.com/

Hammad, N., Eiszler, T., Gazda, R., Cartmell, J., Harpstead, E., & Hammer, J. (2023). V-Light: Leveraging Edge Computing for the Design of Mobile Augmented Reality Games. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3582437.3582456

Hazar, Z. (2018). An Analysis of the Relationship between Digital Game Playing Motivation and Digital Game Addiction among Children. *Asian Journal of Education and Training*, *5*(1). https://doi.org/10.20448/journal.522.2019.51.31.38

Healey, C., & Enns, J. (2012). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, *18*(7). https://doi.org/10.1109/TVCG.2011.127

Hemenover, S. H., & Bowman, N. D. (2018). Video games, emotion, and emotion regulation: expanding the scope. *Annals of the International Communication Association*, *42*(2). https://doi.org/10.1080/23808985.2018.1442239

Ivanina, E. O., Tokmovtseva, A. D., & Akelieva, E. V. (2023). EmoEye: Eye-Tracking and Biometrics Database for Emotion Recognition. *Lurian Journal*, *4*(1). https://doi.org/10.15826/lurian.2023.4.1.1

Jacobwdym. (2018). *Looking at grenades thrown in CS:GO, Kaggle*. Https://Www.Kaggle.Com/Code/Jacobwdym/Looking-at-Grenades-Thrown-in-Cs-Go. https://www.kaggle.com/code/jacobwdym/looking-at-grenades-thrown-in-cs-go

Jeong, J. H., Park, H. J., Yeo, S. H., & Kim, H. (2020). A multimodal analysis combining behavioral experiments and survey-based methods to assess the cognitive effect of video game playing: Good or evil? *Sensors (Switzerland)*, *20*(11). https://doi.org/10.3390/s20113219

Jim Blackhurst. (2011, May 17). *Heatmaps, point clouds and big data in processing*. Https://Web.Archive.Org/Web/20110819001255/Http://Jimblackhurst.Com/Wp/?P=213. https://web.archive.org/web/20110819001255/http://jimblackhurst.com/wp/?p=213

Johari, R., Pekelis, L., & Walsh, D. J. (2015). *Always Valid Inference: Bringing Sequential Analysis to A/B Testing*.

Juliani, A., Berges, V.-P., Teng, E., Cohen, A., Harper, J., Elion, C., Goy, C., Gao, Y., Henry, H., Mattar, M., & Lange, D. (2018). *Unity: A General Platform for Intelligent Agents*.

Junaidi, Julianto, A., Anwar, N., Safrizal, Warnars, H. L. H. S., & Hashimoto, K. (2018). Perfecting a video game with game metrics. *Telkomnika (Telecommunication Computing Electronics and Control)*, *16*(3). https://doi.org/10.12928/

TELKOMNIKA.v16i3.7209

Kabakov, M., Canossa, A., Seif El-Nasr, M., Badler, J. B., Colvin, C. R., Tignor, S., Chen, Z., & Asarsa, K. (2014). A bottom-up method for developing a trait-based model of player behavior. *CHI PLAY 2014 - Proceedings of the 2014 Annual Symposium on Computer-Human Interaction in Play*. https://doi.org/10.1145/2658537.2661320

Kanervisto, A., Kinnunen, T., & Hautamaki, V. (2023). GAN-Aimbots: Using Machine Learning for Cheating in First Person Shooters. *IEEE Transactions on Games*, *15*(4). https://doi.org/10.1109/TG.2022.3173450

Karpouzis, K., & Tsatiris, G. A. (2022). AI in (and for) Games. In *Learning and Analytics in Intelligent Systems* (Vol. 23). https://doi.org/10.1007/978-3-030-76794-5_3

Kaur, D. P., Singh, N. P., & Banerjee, B. (2023). A review of platforms for simulating embodied agents in 3D virtual environments. *Artificial Intelligence Review*, *56*(4). https://doi.org/10.1007/s10462-022-10253-x

Kaye, L. K. (2019). Gaming Classifications and Player Demographics. In *The Oxford Handbook of Cyberpsychology*. https://doi.org/10.1093/oxfordhb/9780198812746.013.1

Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. In *IEEE Computer Graphics and Applications* (Vol. 33, Issue 4). https://doi.org/10.1109/MCG.2013.54

Khalifa, A., Bontrager, P., Earle, S., & Togelius, J. (2020). PCGRL: Procedural Content Generation via Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *16*(1), 95–101. https://doi.org/10.1609/aiide.v16i1.7416

Kinkade, N., Jolla, L., & Lim, K. (2015). DOTA 2 Win Prediction. *University of California*.

Klimas, P. (2019). Current Revenue (Monetisation) Models of Video Gamę Developers. *Journal of Management and Financial Sciences*, *28*. https://doi.org/10.33119/jmfs.2017.28.5

Koivisto, J. M., Havola, S., Engblom, J., Buure, T., Rosqvist, K., & Haavisto, E. (2023). Nursing Students' Scenario Performance: Game Metrics in a Simulation Game. *Nursing Education Perspectives*, *44*(4). https://doi.org/10.1097/01.NEP.0000000000001094

Kotsia, I., Zafeiriou, S., & Fotopoulos, S. (2013). Affective gaming: A comprehensive survey. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/CVPRW.2013.100

Koulaxidis, G., & Xinogalos, S. (2022). Improving Mobile Game Performance with Basic Optimization Techniques in Unity. *Modelling*, *3*(2). https://doi.org/10.3390/modelling3020014

Kozik, A., Machalewski, T., Marek, M., & Ochmann, A. (2021). Mimicking Playstyle by Adapting Parameterized Behavior Trees in RTS Games. *ArXiv Preprint*, 1–8. https://doi.org/10.48550/arXiv.2111.12144

Kriglstein, S. (2019). A taxonomy of visualizations for gameplay data. In *Data Analytics Applications in Gaming and Entertainment*. https://doi.org/10.1201/9780429286490-11

Krstić, N. (2021). Digital playground: Friend or foe to the children? *The European Journal of Applied Economics*, *18*(2). https://doi.org/10.5937/ejae18-29481

Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv:2002.11836*.

Kumaran, V., Rowe, J., Mott, B., & Lester, J. (2023). SCENECRAFT: Automating

Interactive Narrative Scene Generation in Digital Games with Large Language Models. *Proceedings - AAAI Artificial Intelligence and Interactive Digital Entertainment Conference, AIIDE*, *19*(1). https://doi.org/10.1609/aiide.v19i1.27504

Laakkonen, J., Parkkila, J., Jäppinen, P., Ikonen, J., & Seffah, A. (2016). Incorporating Privacy into Digital Game Platform Design: The What, Why, and How. *IEEE Security and Privacy*, *14*(4). https://doi.org/10.1109/MSP.2016.87

Lameman, B. A., El-Nasr, M. S., Drachen, A., Foster, W., Moura, D., & Aghabeigi, B. (2010). User studies - A strategy towards a successful industry-academic relationship. *Future Play 2010: Research, Play, Share - International Academic Conference on the Future of Game Design and Technology*. https://doi.org/10.1145/1920778.1920798

Lee, C. H., & Yoon, H. J. (2017). Medical big data: Promise and challenges. *Kidney Research and Clinical Practice*, *36*(1). https://doi.org/10.23876/j.krcp.2017.36.1.3

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*(3). https://doi.org/10.1016/j.bushor.2017.01.004

Lee, S. J., Jeong, E. J., Lee, D. Y., & Kim, G. M. (2021). Why Do Some Users Become Enticed to Cheating in Competitive Online Games? An Empirical Study of Cheating Focused on Competitive Motivation, Self-Esteem, and Aggression. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.768825

Li, W., Funk, M., Li, Q., & Brombacher, A. (2019). Visualizing event sequence game data to understand player's skill growth through behavior complexity. *Journal of Visualization*, *22*(4), 833–850. https://doi.org/10.1007/s12650-019-00566-5

Lim, C. U., & Harrell, D. F. (2014). Developing social identity models of players from game telemetry data. *Proceedings of the 10th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2014*. https://doi.org/10.1609/aiide.v10i1.12723

Lim, J. Z., Mountstephens, J., & Teo, J. (2022). Eye-Tracking Feature Extraction for Biometric Machine Learning. In *Frontiers in Neurorobotics* (Vol. 15). https://doi.org/10.3389/fnbot.2021.796895

Lin, D., Bezemer, C. P., Zou, Y., & Hassan, A. E. (2019). An empirical study of game reviews on the Steam platform. *Empirical Software Engineering*, *24*(1). https://doi.org/10.1007/s10664-018-9627-4

Lin, L., Liao, X., Jin, H., & Li, P. (2019). Computation Offloading Toward Edge Computing. In *Proceedings of the IEEE* (Vol. 107, Issue 8). https://doi.org/10.1109/JPROC.2019.2922285

Liu, H. X., & Wagner, C. (2023). Proxies to the monthly active user number of Geo AR Mobile games – online search volume as a proposal. *Multimedia Tools and Applications*, *82*(16). https://doi.org/10.1007/s11042-023-14366-5

Liu, J., Snodgrass, S., Khalifa, A., Risi, S., Yannakakis, G. N., & Togelius, J. (2021). Deep learning for procedural content generation. *Neural Computing and Applications*, *33*(1). https://doi.org/10.1007/s00521-020-05383-8

Liu, M., Sun, X., Varshney, M., & Xu, Y. (2019). *Large-Scale Online Experimentation with Quantile Metrics*.

Lop, N. S., Ismail, K., & Mohd Isa, H. (2017). The Implementation of Key Performance Indicators in the Malaysian Private Finance Initiative Projects. *Environment-Behaviour Proceedings Journal*, *2*(5). https://doi.org/10.21834/e-bpj.v2i5.686

López-Gil, J. M., Virgili-Gomá, J., Gil, R., & García, R. (2016). Method for improving EEG based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way. *Frontiers in Computational Neuroscience*, *10*(AUG). https://doi.org/10.3389/fncom.2016.00085

Lynn, J. (2013). Combining Back-End Telemetry Data with Established User Testing Protocols: A Love Story. In *Game Analytics*. https://doi.org/10.1007/978-1-4471-4769-5_22

Makarov, I., Savostyanov, D., Litvyakov, B., & Ignatov, D. I. (2018). Predicting winning team and probabilistic ratings in "Dota 2" and "Counter-strike: Global offensive" video games. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10716 LNCS*. https://doi.org/10.1007/978-3-319-73013-4_17

Mäntymäki, M., Hyrynsalmi, S., & Koskenvoima, A. (2020). How Do Small and Medium-Sized Game Companies Use Analytics? An Attention-Based View of Game Analytics. *Information Systems Frontiers*, *22*(5). https://doi.org/10.1007/s10796-019-09913-1

Martín, M., Jiménez-Martín, A., Mateos, A., & Hernández, J. Z. (2021). Improving a/b testing on the basis of possibilistic reward methods: A numerical analysis. *Symmetry*, *13*(11). https://doi.org/10.3390/sym13112175

Matsui, A., Sapienza, A., & Ferrara, E. (2020). Does Streaming Esports Affect Players' Behavior and Performance? *Games and Culture*, *15*(1). https://doi.org/10.1177/1555412019838095

McDaniel, R., & Yampolskiy, R. V. (2012). Development of embedded CAPTCHA elements for bot prevention in Fischer random chess. *International Journal of Computer Games Technology*. https://doi.org/10.1155/2012/178578

Melhart, D., Togelius, J., Mikkelsen, B., Holmgard, C., & Yannakakis, G. N. (2024). The Ethics of AI in Games. *IEEE Transactions on Affective Computing*, *15*(1). https://doi.org/10.1109/TAFFC.2023.3276425

Merrick, K. (2008). Modeling motivation for adaptive nonplayer characters in dynamic computer game worlds. *Computers in Entertainment*, *5*(4). https://doi.org/10.1145/1324198.1324203

Meyer, B. C., & Bishop, D. S. (2022). A lesson in Tableau dashboard design: Playing the beer game with a real-time data connection. *Decision Sciences Journal of Innovative Education*, *20*(4). https://doi.org/10.1111/dsji.12264

Micallef, D., Schivinski, B., Brennan, L., Parker, L., & Jackson, M. (2024). "What Are You Eating?" Is the Influence of Fortnite Streamers Expanding Beyond the Game? *Journal of Electronic Gaming and Esports*, *2*(1). https://doi.org/10.1123/jege.2023-0033

Mirza-Babaei, P., Nacke, L. E., Wallner, G., & McAllister, G. (2014). Unified visualization of quantitative and qualitative playtesting data. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2559206.2581224

Moura, D., El-Nasr, M. S., & Shaw, C. D. (2011). Visualizing and understanding players' behavior in video games: Discovering patterns and supporting aggregation and comparison. *ACM SIGGRAPH 2011 Game Papers, SIGGRAPH'11*. https://doi.org/10.1145/2037692.2037695

Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis. In *Big Data and Cognitive Computing* (Vol. 4, Issue 2).

https://doi.org/10.3390/bdcc4020004

Na, J. C., Kim, E. J., & Kim, J. Y. (2022). Big data analysis of the impact of COVID-19 on digital game industrial sustainability in South Korea. *PLoS ONE*, *17*(12 December). https://doi.org/10.1371/journal.pone.0278467

Nassi, J. J., & Callaway, E. M. (2009). Parallel processing strategies of the primate visual system. In *Nature Reviews Neuroscience* (Vol. 10, Issue 5). https://doi.org/10.1038/nrn2619

Paajala, I., Nyyssölä, J., Mattila, J., & Karppinen, P. (2022). Users' Perceptions of Key Blockchain Features in Games. *Future Internet*, *14*(11). https://doi.org/10.3390/fi14110321

Paduraru, C., Cristea, R., & Stefanescu, A. (2022). Enhancing the security of gaming transactions using blockchain technology. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3551349.3560504

Pfau, J., Liapis, A., Volkmar, G., Yannakakis, G. N., & Malaka, R. (2020). Dungeons Replicants: Automated Game Balancing via Deep Player Behavior Modeling. *IEEE Conference on Computatonal Intelligence and Games, CIG*, *2020-August*. https://doi.org/10.1109/CoG47356.2020.9231958

Pirker, J., Griesmayr, S., Drachen, A., & Sifa, R. (2016). How playstyles evolve: Progression analysis and profiling in Just Cause 2. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9926 LNCS*. https://doi.org/10.1007/978-3-319-46100-7_8

Plans, D., & Morelli, D. (2012). Experience-driven procedural music generation for games. *IEEE Transactions on Computational Intelligence and AI in Games*, *4*(3). https://doi.org/10.1109/TCIAIG.2012.2212899

Quin, F., Weyns, D., Galster, M., & Silva, C. C. (2023). *A/B Testing: A Systematic Literature Review*.

Radhakrishnan, K., Baranowski, T., O'Hair, M., Fournier, C. A., Spranger, C. B., & Kim, M. T. (2020). Personalizing sensor-controlled digital gaming to self-management needs of older adults with heart failure: A qualitative study. *Games for Health Journal*, *9*(4). https://doi.org/10.1089/g4h.2019.0222

Rego, I. B., Coelho, S., Semedo, P. M., Cavaco-Silva, J., Teixeira, L., Sousa, S., Reis, J., Dinis, R., Schmitt, F., Afonso, N., Fougo, J. L., Pavão, F., Baptista Leite, R., & Costa, L. (2023). 360 Health Analysis (H360)—A Comparison of Key Performance Indicators in Breast Cancer Management across Health Institution Settings in Portugal. *Current Oncology*, *30*(7). https://doi.org/10.3390/curroncol30070451

Renshaw, T., Stevens, R., & Denton, P. D. (2009). Towards understanding engagement in games: An eye-tracking study. *On the Horizon*, *17*(4). https://doi.org/10.1108/10748120910998425

Reynolds, J. (2019). Gambling on big data: Designing risk in social casino games. In *European Journal of Risk Regulation* (Vol. 10, Issue 1). https://doi.org/10.1017/err.2019.18

Roohi, S., Relas, A., Takatalo, J., Heiskanen, H., & Hämäläinen, P. (2020). Predicting Game Difficulty and Churn without Players. *CHI PLAY 2020 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. https://doi.org/10.1145/3410404.3414235

Ruiperez-Valiente, J. A., Gomez, M. J., Martinez, P. A., & Kim, Y. J. (2021). Ideating and Developing a Visualization Dashboard to Support Teachers Using Educational

Games in the Classroom. *IEEE Access*, *9*, 83467–83481. https://doi.org/10.1109/ACCESS.2021.3086703

Russell, N. C., Reidenberg, J. R., & Moon, S. (2018). Privacy in gaming. Fordham Intell. Prop. Media & Ent. LJ. doi:10.2139/ssrn.3147068

Sahoo, S. (2022). Big data analytics in manufacturing: a bibliometric analysis of research in the field of business management. In *International Journal of Production Research* (Vol. 60, Issue 22). https://doi.org/10.1080/00207543.2021.1919333

Sapienza, A., Peng, H., & Ferrara, E. (2017). Performance dynamics and success in online games. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2017-November*. https://doi.org/10.1109/ICDMW.2017.124

Schiller, M. H., Wallner, G., Schinnerl, C., Calvo, A. M., Pirker, J., Sifa, R., & Drachen, A. (2019). Inside the group: Investigating social structures in player groups and their influence on activity. *IEEE Transactions on Games*, *11*(4). https://doi.org/10.1109/TG.2018.2858024

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2023). Artificial Intelligence Governance For Businesses. *Information Systems Management*, *40*(3). https://doi.org/10.1080/10580530.2022.2085825

Seif El-Nasr, M. (2019). Developing games that capture and engage users. *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion, ICSE-Companion 2019*. https://doi.org/10.1109/ICSE-Companion.2019.00025

Seif El-Nasr, M., Gagné, A., Moura, D., & Aghabeigi, B. (2013). Visual Analytics Tools – A Lens into Player's Temporal Progression and Behavior. In *Game Analytics*. https://doi.org/10.1007/978-1-4471-4769-5_19

Seif El-Nasr, M., & Kleinman, E. (2020). Data-Driven Game Development: Ethical Considerations. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3402942.3402964

Sifa, R., Bauckhage, C., & Drachen, A. (2014). The Playtime Principle: Large-scale cross-games interest modeling. *2014 IEEE Conference on Computational Intelligence and Games*, 1–8. https://doi.org/10.1109/CIG.2014.6932906

Sifa, R., Drachen, A., & Bauckhage, C. (2018). Profiling in Games: Understanding Behavior from Telemetry. In K. Lakkaraju, G. Sukthankar, & R. T. Wigand (Eds.), *Social Interactions in Virtual Worlds: An Interdisciplinary Perspective* (pp. 337–374). Cambridge University Press. https://doi.org/DOI: 10.1017/9781316422823.014

Sifa, R., Drachen, A., Bauckhage, C., Thurau, C., & Canossa, A. (2013). Behavior evolution in Tomb Raider Underworld. *IEEE Conference on Computatonal Intelligence and Games, CIG*. https://doi.org/10.1109/CIG.2013.6633637

Silva, C. (2016). Data visualization methods. *Journal of the Acoustical Society of America*, *140*(4_Supplement). https://doi.org/10.1121/1.4970803

Simoes, R., & Gomes, A. (2023). Assessing the engagement of players in analog board games through biometric monitoring. *Iberian Conference on Information Systems and Technologies, CISTI*, *2023-June*. https://doi.org/10.23919/CISTI58278.2023.10211662

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., Tripoliti, E., Marias, K., Fotiadis, D. I., & Tsiknakis, M. (2023). Review of Eye Tracking Metrics Involved in Emotional and Cognitive Processes. In *IEEE Reviews in Biomedical Engineering* (Vol. 16). https://doi.org/10.1109/RBME.2021.3066072

Smeddinck, J., Gerling, K. M., & Tiemkeo, S. (2013). Visual complexity, player experience, performance and physical exertion in motion-based games for older adults. *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2013*. https://doi.org/10.1145/2513383.2517029

Smerdov, A., Somov, A., Burnaev, E., Zhou, B., & Lukowicz, P. (2021). Detecting Video Game Player Burnout with the Use of Sensor Data and Machine Learning. *IEEE Internet of Things Journal*, *8*(22). https://doi.org/10.1109/JIOT.2021.3074740

Soares, R., Sarmanho, E., Miura, M., Silva, T., & Castanho, C. (2016). Biofeedback Sensors in Game Telemetry Research. *Xv Simpósio Brasileiro de Jogos e Entretenimento Digital*.

Strubberg, B. C., Elliott, T. J., Pumroy, E. P., & Shaffer, A. E. (2020). Measuring Fun. *Loading*, *13*(21). https://doi.org/10.7202/1071448ar

Su, Y., Backlund, P., & Engström, H. (2022). Data-driven method for mobile game publishing revenue forecast. *Service Oriented Computing and Applications*, *16*(1). https://doi.org/10.1007/s11761-021-00332-2

Summerville, A., Snodgrass, S., Guzdial, M., Holmgård, C., Hoover, A. K., Isaksen, A., Nealen, A., & Togelius, J. (2017). Procedural Content Generation via Machine Learning (PCGML). *Arxiv Preprint*.

Sun, Y., Li, Z., Fang, K., Lee, C. H., & Asadipour, A. (2023). Language as Reality: A Co-creative Storytelling Game Experience in 1001 Nights Using Generative AI. *Proceedings - AAAI Artificial Intelligence and Interactive Digital Entertainment Conference, AIIDE*, *19*(1). https://doi.org/10.1609/aiide.v19i1.27539

Taesiri, M. R., Macklon, F., Wang, Y., Shen, H., & Bezemer, C.-P. (2022). Large Language Models are Pretty Good Zero-Shot Video Game Bug Detectors. *Arxiv Preprint*.

Tarnowski, P., Kołodziej, M., Majkowski, A., & Rak, R. J. (2020). Eye-Tracking Analysis for Emotion Recognition. *Computational Intelligence and Neuroscience*, *2020*. https://doi.org/10.1155/2020/2909267

Trepte, S., & Reinecke, L. (2011). The pleasures of success: Game-related efficacy experiences as a mediator between player performance and game enjoyment. *Cyberpsychology, Behavior, and Social Networking*, *14*(9). https://doi.org/10.1089/cyber.2010.0358

Van Linden, R. Der, Lopes, R., & Bidarra, R. (2013). Designing procedurally generated levels. *AAAI Workshop - Technical Report*, *WS-13-20*. https://doi.org/10.1609/aiide.v9i3.12592

Vazquez, J., Abdelrahman, S., Wasden, C., Jardine, S., Judd, C., Davis, M., & Facelli, J. C. (2022). Using Biometric Data to Measure and Predict Emotional Engagement of Video Games. *BioRxiv Preprint*. https://doi.org/10.1101/2022.02.28.482337

Vicencio-Moreira, R., Mandryk, R. L., & Gutwin, C. (2015). Now you can compete with anyone: Balancing players of different skill levels in a first-person shooter game. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*. https://doi.org/10.1145/2702123.2702242

Viljanen, M., Airola, A., Heikkonen, J., & Pahikkala, T. (2017). A/B-test of retention and monetization using the cox model. *Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2017*. https://doi.org/10.1609/aiide.v13i1.12941

Wallner, G., & Drachen, A. (2023). Beyond the Game: Charting the Future of Game Data Science. *Games: Research and Practice*, *1*(1). https://doi.org/10.1145/3582933

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1). https://doi.org/10.1109/TKDE.2013.109

Yampray, K., & Inchamnan, W. (2019). A method to visualization data collection by using gamification. *International Conference on ICT and Knowledge Engineering*, *2019-November*. https://doi.org/10.1109/ICTKE47035.2019.8966880

Yang, W., Huang, T., Zeng, J., Chen, L., Mishra, S., & Youjian, L. (2019). Utilizing Players' Playtime Records for Churn Prediction: Mining Playtime Regularity. *IEEE Transactions on Games*, *14*(2), 153–160.

Yu, J., Ma, W., Moon, J., & Denham, A. R. (2022). Developing a Stealth Assessment System Using a Continuous Conjunctive Model. *Journal of Learning Analytics*, *9*(3). https://doi.org/10.18608/jla.2022.7639

Zhou, L. (2023). An Introduction to Data Visualization. *Highlights in Science, Engineering and Technology*, *31*. https://doi.org/10.54097/hset.v31i.4813

Zhou, Z., & Guzdial, M. (2021). Toward Co-creative Dungeon Generation via Transfer Learning. *The 16th International Conference on the Foundations of Digital Games (FDG) 2021*, 1–9. https://doi.org/10.1145/3472538.3472601

Zhu, J., & Ontañón, S. (2020). Player-Centered AI for Automatic Game Personalization: Open Problems. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3402942.3402951

About The Authors

Murat ATASOY, is working in the distance education and digital transformation units at Trabzon University with the title of Research Assistant Doctor. He completed his PhD in the field of Instructional Technologies on the development of 3D Avatar TİD3B, which converts Turkish texts into Turkish Sign Language. In his master's thesis, he worked on software development for automatic material development for teachers of the hearing impaired. He worked as a lecturer and administrative staff at Recep Tayyip Erdoğan University between 2006-2013. He won various awards at TEKNOFEST, took part in many international scientific publications, and worked as a fellow researcher in different TUBITAK and BAP projects. His research areas are instructional technologies, the use of technology in special education, avatars, 2D-3D game development, and virtual environments.
E-mail: murat.atasoy@trabzon.edu.tr, **ORCID:** 0000-0001-6589-0161

Adil YILDIZ, is working in the big data and digital transformation units at Trabzon University with the title of Lecturer Doctor. He completed his PhD in Educational Technologies in 2022. His research areas are Open and distance learning, distance education design and management, educational technologies, and synchronous software platforms. He has won various awards at TEKNOFEST, taken part in many international scientific publications, and worked as a scholar and researcher in different TUBITAK and BAP projects.
E-mail: adilyildiz@trabzon.edu.tr, **ORCID:** 0000-0002-7383-3885

Lokman ŞILBIR, received his Ph.D. in Computer Education and Instructional Technology from Karadeniz Technical University. He currently serves as an Associate Professor in the Computer Technologies Department at Çarşıbaşı Vocational School of Trabzon University. He has participated in many international scientific publications and has worked as a scholar and researcher in different TUBITAK and BAP projects. His research interests include material development, human-computer interaction,

programming, and distance education.

E-mail: lokmansilbir@trabzon.edu.tr, **ORCID:** 0000-0003-3655-2512

Ekrem BAHÇEKAPILI, received a Ph.D. degree from the Department of Computer Education and Instructional Technology at Atatürk University. He works as an Associate Professor in the Department of Management Information Systems of the Faculty of Economics & Administrative Sciences of the Karadeniz Technical University. His research interests include Management Information Systems, Digital Transformation, Human-Computer Interaction, and Technology Acceptance.
E-mail: ekrem.bahcekapili@ktu.edu.tr, **ORCID:** 0000-0002-7538-1712

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 4%.

Artificial Intelligence-Powered Data Analytics against Botnet Attacks: Threat Detection and Ethical Considerations

Ramazan KOCAOĞLU

Ostim Technical University

To Cite This Chapter

Kocaoğlu, R. (2024). Artificial Intelligence-Powered Data Analytics against Botnet Attacks: Threat Detection and Ethical Considerations. In M. Hanefi Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 86-96). ISRES Publishing.

Introduction

In today's digital world, botnet attacks pose serious threats to both individuals and organizations. Botnets allow cyber attackers to carry out large-scale attacks by capturing hundreds of thousands or even millions of devices under a network. These attacks usually involve destructive activities such as DDoS (Distributed Denial of Service), phishing, and spam, causing both financial and operational losses. The ever-increasing size of botnet attacks has necessitated the development of new and powerful methods that can effectively detect and block these threats.

Data analytics and artificial intelligence have the potential to revolutionize the security world in order to protect against botnet attacks (Owen et al., 2022) Data analytics enables tracing and predicting attacks by extracting meaningful patterns from the large amounts of data generated by botnet networks. In addition, artificial intelligence-based methods offer promising solutions to detect botnet attacks with techniques such as behavioral analysis, anomaly detection, and machine learning models. The ability of artificial intelligence to learn over large data sets makes it possible for this technology to adapt to constantly evolving and changing attack techniques.

However, the use of artificial intelligence against botnet attacks raises a number of ethical and technical issues. In particular, issues such as data privacy, the ethical acceptability of artificial intelligence models, and the potential damage that false positives can cause are critical areas that need to be carefully considered when developing such solutions (Stahl, 2021). In addition, how effective existing AI solutions can remain against evolving attack methods is another important element that should be discussed in long-term security strategies.

In this chapter, how AI-supported data analytics methods can be used against botnet attacks, the effectiveness of these technologies in security, the technical and ethical challenges encountered, and future development areas will be evaluated in detail. At the end of the chapter, in light of the solutions offered by artificial intelligence against botnet attacks, recommendations will be presented for future directions and the development of healthier solutions in this field.

Botnet Attacks and Threat Detection with Data Analytics

Botnet attacks, an important part of cyber security threats, pose a great risk in today's digital world. Botnets are a distributed network consisting of a large number of devices under the control of attackers and are generally used for purposes such as DDoS attacks,

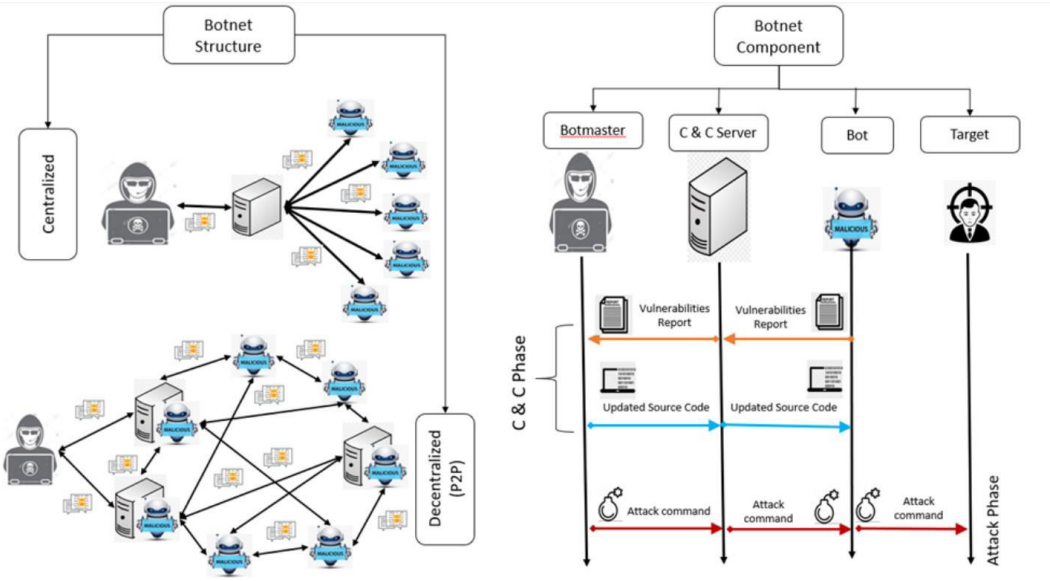
phishing, spamming, and malware propagation. Studies show that botnet attacks have increased rapidly in recent years and have reached much more dangerous dimensions (Ali et al., 2020).

A widely known instance of a botnet attack was the Mirai botnet incident in 2016 (Kambourakis et al., 2017). The Mirai botmaster performed reconnaissance on vulnerable Internet of Things devices connected to the target networks and subsequently deployed the malware against unprotected ports. As a result, the Mirai botnet infected over 600,000 IoT devices, utilizing them to mount a massive DDoS attack that disrupted access to popular websites like Netflix, Twitter, GitHub, and Reddit (Salim et al., 2019). The growing prevalence and sophistication of botnet attacks have necessitated the development of robust and adaptive security measures.

Botnets are large networks of a large number of devices that have been compromised by malware. These devices operate in a coordinated manner through a central command and control (C&C) center under the control of the attackers. The Bootnet structure is given in **Figure 1** (Ibrahim et al., 2022).

Figure 1

The Botnet Structure and Botnet Component



The operating structure and propagation mechanisms of botnets usually consist of the following components:

> *C&C Centre:* The main management center that allows attackers to command the entire bot network.
> *Bots:* Devices connected to the C&C center (computers, IoT devices, etc.) form the main components of the botnet.
> *Command Chain and Propagation:* Botnets attack their targets through commands. They are often spread through social engineering techniques such as malware emails or fake websites.

Botnets can be organized in different ways in terms of propagation and attack diversity. For instance, "Peer-to-Peer" botnets enable bots to communicate directly with one another without requiring a centralized command system, which makes them challenging to detect. These decentralized botnets are frequently employed in a variety of malicious activities, such as DDoS attacks, spamming, phishing, and malware distribution, posing

a significant threat to digital security. Unlike traditional botnets with a centralized command and control structure, peer-to-peer botnets are more resilient and adaptive, as they can continue to function even if individual nodes are taken down. This distributed nature makes them particularly difficult to disrupt and monitor, further exacerbating the security challenges they pose.

Data analytics plays a critical role in security strategies to detect and block botnet attacks (Xing et al., 2021). Botnets leave a certain trace by constantly exhibiting abnormal behavior in network traffic, and data analytics can extract these traces from large data sets. The contributions of data analytics in detecting botnet attacks are as follows :

Anomaly Detection: Botnet attacks often generate anomalous traffic patterns. Data analytics algorithms can identify such anomalies and generate alarms for security systems. For example, activities such as excessive connection requests or unusual data transfer can be quickly detected with data analytics tools (Ahmad et al., 2022).
Behavioral Analysis: Data analytics can predict the attack activity of botnets by analyzing user and device behavior. Botnets often exhibit specific communication patterns, which can be identified through deep inspection of network traffic and device logs. Early warning systems can be established by performing communication detection analysis of bot patterns (Analysis of Botnet Attack Communication Pattern Behavior on Computer Networks, 2022).
Pattern Recognition and Machine Learning: In defense systems against botnet attacks, attack patterns are determined with machine learning models (Yang et al., 2022). These models can detect new threats faster by learning from past botnet attacks. Especially in DDoS attacks, botnet-based attacks can be monitored more effectively with machine learning algorithms.

These advantages of data analytics strengthen digital security by detecting botnets and reducing the potential for attacks. However, processing large amounts of data and performing real-time analyses can pose both technical and ethical challenges. Nevertheless, data analytics remains an essential component of modern cybersecurity strategies against botnets.

The Effectiveness of Artificial Intelligence Methods in Botnet Detection

Artificial intelligence-based data analytics offers a powerful solution for the detection of botnet attacks, especially through the use of machine learning and deep learning models. In this chapter, the main artificial intelligence methods used in the detection of botnet threats and the effectiveness of these methods will be discussed.

Machine Learning and Botnet Detection

Machine learning provides effective results with classification and clustering techniques commonly used to detect botnet attacks. These techniques allow us to distinguish and categorize botnet activities thanks to algorithms trained on historical data (Yang et al., 2022).

Classification Algorithms: Classification algorithms are used to distinguish botnet traffic from normal network traffic. Classifiers classify the samples in the data set into certain classes in order to predict botnet-based attacks. For example, methods such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees are frequently preferred to detect DDoS attacks. However, classification models have a margin of error such as false positives and false negatives, and the accuracy rate depends on the data quality and the training of the model.

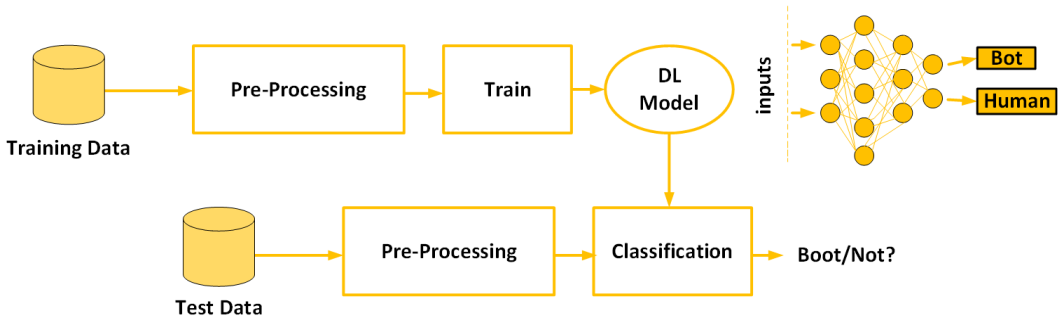
Clustering Techniques: Clustering methods aim to create anomalous clusters in the data set by bringing together similar data points. Clustering, one of the unsupervised learning techniques, is effective in distinguishing botnet activities in previously unlabelled data sets. Methods such as K-Means and Hierarchical Clustering identify the points where botnet activities are concentrated and compare them with normal traffic. However, it can be challenging to correctly identify clusters based on data density and attack type.

Machine learning models enable faster detection of botnet attacks by processing large amounts of data. However, the lack of sufficient data labels, the complexity of attacks, and rapidly changing attack techniques require machine learning-based models to be updated.

Detection with Deep Learning Models

Deep learning is an important method in botnet detection with its capacity to process large amounts of data and its ability to detect complex attack patterns (Yerima et al., 2021). In particular, neural network-based deep learning models perform effectively in large datasets. Figure 2 shows the general steps for using deep learning to detect bots. First, the data is cleaned up by labeling, expanding, and extracting features. This prepared data is then used to train a deep-learning model. Finally, the trained model is used to classify whether users in new data are human or bots.

Figure 2
A Generalized Workflow Diagram DL-Based Bot Detection



Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN): ANN and RNN are effective in identifying relationships between events by processing temporal data of botnet attacks. RNNs analyze how botnet behavior can be related to past events, especially in time series data.
Convolutional Neural Networks (CNN): CNN is a powerful method for identifying botnet attack traces in network traffic data. CNN, which is often used for image and pattern recognition, can extract data from network traffic to distinguish the specific traffic of botnet attacks.

Although deep learning models have high accuracy rates in botnet detection, they require a large data processing capacity. Preparation, modeling and training of datasets require large amounts of time and computational power. Furthermore, deep learning models, while offering high accuracy, can be of limited use in real-time applications due to excess complexity.

Anomaly Detection Methods

Anomaly detection is based on detecting botnet attacks through unusual network activities. Identifying activity that deviates from normal network behavior is an effective

method for botnet detection. This method can be achieved by various algorithms:

Statistical Methods: For the detection of botnet attacks, statistical deviations from normal are analyzed. For example, abnormal behavior, such as a device making more connections than expected, can be determined by statistical anomaly detection (Ashraf et al., 2021).

Machine Learning-Based Anomaly Detection: Anomaly detection methods supported by classification algorithms are used to find anomalous patterns of suspicious activity in data sets. Isolation Forest is one of the common anomaly detection methods used to detect botnet attacks (Borges et al., 2023). This algorithm signals botnet activities by extracting unusual points in the data set.

Time-Series Based Anomaly Analysis: The fact that botnet attacks lead to continuous traffic makes it easy to detect anomalies with time-series analysis. Time series analysis methods can identify attack symptoms by capturing the behavioral patterns of botnets over time(Borges et al., 2023).

Anomaly detection is effective in predicting botnet attacks; however, a high false positive rate may lead security teams to receive unnecessary alarms. Therefore, anomaly detection models should be fine-tuned according to the data set and network traffic characteristics.

Ethical and Social Aspects of Artificial Intelligence Methods in Botnet Detection

Artificial intelligence-based security solutions play an important role in the detection and prevention of botnet attacks. However, it is necessary to consider the ethical and social dimensions of the use of these systems. In this section, the ethical and social issues of AI-based botnet detection systems, such as data privacy, fairness, and regulations, will be discussed.

Data Privacy and Security Concerns

AI-based security solutions detect botnet threats by analyzing large amounts of data, raising concerns about personal data privacy and data security. AI algorithms used for botnet detection monitor users' online activities, network traffic patterns, and other sensitive data. Privacy and security issues come to the forefront in these data analysis processes:

Data Privacy Breaches: Data collected for the training of artificial intelligence models may contain users' personal information. Threats to user privacy arise, particularly when analysing sensitive data such as network traffic. Storing personal data in AI systems increases the risk of misuse through data breaches and cyber-attacks.

Vulnerabilities: The security of AI systems must be resilient against botnet attacks. However, the vulnerabilities of artificial intelligence models themselves can reduce the reliability of these systems. For example, the effectiveness of security solutions can be jeopardized if attackers infiltrate the model training process and manipulate the model (adversarial attacks).

These concerns require AI-based botnet detection systems to adopt a more transparent and privacy-oriented approach to data collection and storage. If data security is not ensured, ethical commitments to protect users' privacy may be violated (Dhirani et al., 2023).

False Alarms and Justice Issues

The false positive rates of AI algorithms in botnet detection pose ethical and social challenges. In particular, false positive alarms can lead to innocent users being unjustly associated with botnet activity and may create injustice.

The Impact of False Positives: Artificial intelligence algorithms used in botnet detection may sometimes mistakenly recognize normal user activity as a botnet

attack. This can cause innocent users to be subjected to unnecessary security reviews. For example, network activity of a business that generates heavy network traffic may be incorrectly interpreted as botnet activity. A high number of false positives can cause security teams to spend unnecessary time and users to be unfairly blamed.

Fairness Issues and Model Biases: In order for AI-based security systems to make fair decisions, imbalances in data sets should be considered during the training of models. For example, if user data from certain geographical areas or demographic groups is analyzed more, these groups may be targeted more than others. This may be contrary to principles of social justice and may result in users being unfairly discriminated against by certain groups.

In order to fulfill ethical commitments in terms of fairness and accuracy, it is important to balance the precision and accuracy of AI algorithms, reducing false positives and ensuring the impartiality of algorithms.

Regulations and Policies for Artificial Intelligence Applications

Various regulations and policies are required to fulfill ethical responsibilities during the development and implementation of AI-based botnet detection systems. Such regulations encourage the use of artificial intelligence in the security sector in accordance with ethical principles.

Existing regulations: Today, general data protection regulations, such as GDPR, are in place to protect data privacy. Such regulations oblige artificial intelligence systems to protect user privacy in data collection and analysis processes. However, there is no direct regulation for specific security applications such as botnet detection.

Proposed Regulations: New policies are proposed to increase the transparency and fairness of AI applications. For example, algorithmic transparency allows users to understand how their data is used and what decision processes it influences. It is also recommended to develop standardized testing processes to reduce false positive rates and protocols to ensure algorithm reliability.

Implementation of Ethical Principles and Policies: To ensure ethical commitments in the security applications of AI, policies should cover not only technical standards but also values such as transparency, fairness, and trustworthiness. This aims to ensure that AI-based systems provide effective protection against botnet attacks while protecting user rights.

The effective implementation of these regulations and policies will contribute to the ethical development and reliability of AI-based botnet detection solutions.

Technical Challenges in Artificial Intelligence-Based Botnet Detection

While AI-supported botnet detection systems offer powerful solutions in the field of cyber security, they also face various technical challenges (Zhang et al., 2021). These challenges can affect the accuracy and effectiveness of the systems and make it difficult to develop real-time and adaptive solutions to botnet attacks. In this section, we will discuss the main technical obstacles faced by artificial intelligence applications in botnet detection.

Data Quality and Labelling Problems

In order for artificial intelligence models to accurately detect botnet attacks, high-quality and comprehensive training data is required. However, data quality and data labeling problems are important technical obstacles that directly affect model performance.

Missing or Low-Quality Data: Missing or poor-quality data for botnet attacks makes it difficult for the model to accurately learn attack types. Inadequate data can reduce the reliability of the model, causing it to fail to correctly identify attack patterns. For example, failure to obtain complete network traffic data may lead to missing salient characteristics of botnet activity.

Labelling Problems: Machine learning models need to be trained with correctly labeled data. However, in cybersecurity incidents such as botnet attacks, accurate labeling can be difficult because attacks are complex and constantly evolving. Incorrect or incompletely labeled data can increase the false positive and false negative rates of the model and lead to erroneous results.

Overcoming data quality and labeling problems necessitates the improvement of data collection processes for botnet detection and the use of more advanced labeling methods.

Real-time Detection and Performance Problems

Detecting botnet attacks in real-time is a major technical challenge, especially considering the need to deal with large-scale data sets and high-speed network traffic. This puts pressure on the processing power and performance of AI models.

Large-Scale Data Processing: Most botnet attacks can be hidden in large chunks of data. Especially in large-scale attacks such as DDoS, it is necessary for systems to analyze large amounts of data in real time. However, analyzing large data sets on the fly requires high processing power and memory, which can slow or limit system performance.

Real-time Analysis Requirement: Botnet attacks may require an immediate response; otherwise, the damage can grow rapidly. However, artificial intelligence models may experience delays while performing complex calculations on large amounts of data. In particular, deep learning models may encounter performance problems in real-time analyses due to computational intensity.

In order to overcome such performance problems, the development of lightweight and fast-running models or the use of methods such as parallel processing can facilitate real-time results in AI-based botnet detection.

Adaptation and Update Requirements

Botnet attacks are constantly evolving, developing new techniques and increasing their ability to circumvent security measures. Therefore, AI-based systems need to be regularly updated to adapt to the current state of the attacks.

Adaptation to Evolving Attack Techniques: New variations of botnet attacks can differ from traditional attack techniques. AI models need to be regularly retrained or updated to identify these changes and detect new threats. However, updating the models in this way requires continuous reconfiguration of the data collection and processing processes.

Need for Dynamic Updating: Artificial intelligence models must be dynamically updated to learn about changes in attack types. However, this process is costly in terms of both time and processing power and may affect the uptime of systems. In addition, the constant change in attacks necessitates periodic retraining of artificial intelligence models in order to maintain their accuracy in the long term.

Adaptation and updating needs constitute an important requirement for AI-based botnet detection systems to keep up with attack dynamics. In order to overcome these problems, innovative solutions such as online learning methods and continuously updated model structures are being studied.

Future Perspectives: Artificial Intelligence and the Evolution of Botnet Attacks

Emerging developments in artificial intelligence technologies may offer powerful solutions for detecting and preventing botnet attacks. In particular, advancements in algorithms and hardware can enhance the impact of AI systems in the security domain.

Advanced Deep Learning Models: Researchers are developing deeper and larger-scale neural networks that demonstrate high accuracy in botnet detection. For instance, Transformer-based models and graph-based neural networks have the potential to more precisely recognize botnet behavior patterns. Such advanced algorithms can provide more accurate detections by effectively processing extensive datasets.

Autonomous Learning Systems: AI can adapt to evolving threats through autonomous learning systems. Specifically, AI models supported by techniques like Reinforcement Learning can generate more effective solutions against botnet attacks by self-learning.

Quantum Computing and AI: Developments in quantum computing could revolutionize AI-based botnet detection in the future. Quantum computing can quickly analyze much larger datasets, enabling real-time threat detection. This capability can help more effectively block large-scale botnet attacks.

In the coming years, new artificial intelligence-supported developments are expected in areas such as autonomous systems in the detection of botnet attacks, threat intelligence, and real-time analysis. These developments can provide both more integrated and more effective solutions in the field of security.

Autonomous Defence Systems: In the future, AI-enabled botnet detection systems will be able to operate autonomously without human intervention. These systems can provide faster and more effective defense by recognizing and responding to attacks in real-time. For example, artificial intelligence algorithms integrated with systems such as firewalls or IPS (Intrusion Prevention Systems) can automatically trigger and block the attack at the time of attack.

Integration with Threat Intelligence Systems: AI-powered botnet detection will become part of a broader security network in the future by integrating with threat intelligence systems. Such integrated systems can quickly recognize the first signs of botnet attacks by analyzing threat information on a global scale and transferring threat information to other security elements.

Real-Time and Advanced Algorithms: By making artificial intelligence algorithms faster and more effective, the real-time analysis capacity will increase. Thus, botnet attacks can be detected as soon as they start and can be stopped before the attack spreads. Real-time detection will provide higher accuracy rates, especially in large-scale and complex attacks.

Ethics and Safety Best Practices

Determining ethical and technical best practices in the use of artificial intelligence against botnet attacks is of great importance for ensuring security and protecting ethical rules. These practices ensure the effective and responsible use of AI-based security solutions.

Transparency and Accountability: It is important to inform users about what data AI systems use and what kind of analysis they perform when detecting botnets. Transparency has become an ethical imperative for both the protection of user privacy and the reliability of the system. Accountable AI applications also provide an important safety net to protect users as a result of faulty decisions.

Reducing False Positives: Reducing the false positive rates of AI systems in

botnet detection is critical to improving the user experience. Unjustifiably associating users with botnet activity due to false positives can lead to trust issues. Best practices should include verification techniques and multi-layer verification mechanisms to minimize false positive rates.

Data Privacy and Security: User data must be protected in AI-based security solutions. Best practices such as data anonymization, data minimization, and processing only the data that is necessary should be adopted to reduce the risk of privacy breaches. In addition, security measures should be increased to ensure that user data is stored securely and not shared.

The implementation of these best practices will enable AI-based security systems to develop within the framework of ethical rules and increase user security. In the future, these principles will contribute to the evolution of artificial intelligence systems to benefit society by creating a more effective and reliable defense mechanism against botnet attacks.

Conclusion

Artificial intelligence-supported data analytics stands out with its various advantages in detecting and preventing botnet attacks. Its ability to quickly and accurately analyze large-scale data sets allows it to quickly detect a large number of botnet activities. In addition, methods such as machine learning and deep learning offer strong performance in recognizing patterns specific to botnet attacks and supporting automatic threat detection processes based on these patterns.

However, AI-supported botnet detection solutions have some limitations. First of all, high-quality and comprehensive training data is needed for AI models to be successful. Lack of data, low-quality data, or mislabelled data may negatively affect model accuracy. In addition, as botnet attacks constantly evolve and develop new ways against security systems, it is imperative that artificial intelligence systems adapt quickly to these changes. This adaptation need can be costly in terms of time and resources, as it requires constantly updated and evolving structures.

There are important opportunities for the development of artificial intelligence-based systems against botnet attacks. In particular, advances in areas such as advanced learning algorithms, autonomous threat detection systems, and real-time analysis capabilities will increase the effectiveness in this area and make security stronger.

The evolution of botnet attacks and the future of artificial intelligence-supported solutions against these attacks offer new perspectives in the field of security. In the future, it is expected to develop systems that are more autonomous, flexible, and capable of real-time analyses. Furthermore, with the integration of quantum computing and advanced deep learning algorithms into this field, artificial intelligence models with much more powerful data processing capacity may emerge.

With the advancement of the adaptive learning capabilities of artificial intelligence, systems used in botnet detection will become more resilient to dynamic attacks. In addition, ethical and security best practices are expected to gain more importance in this field. Issues such as data privacy, false positives, and fairness will be among the most important components of AI-supported security solutions in the future.

In conclusion, there are both technical and ethical development opportunities in the field of AI-based botnet detection. The development of technological innovations and applications in this field will continue to provide stronger, reliable, and ethical solutions in the field of cyber security.

References

Ahmad, S., Jha, S., Alam, A., Alharbi, M., & Nazeer, J. (2022, May 12). Analysis of Intrusion Detection Approaches for Network Traffic Anomalies with Comparative Analysis on Botnets (2008–2020). Hindawi Publishing Corporation, 2022(undefined), 1-11. https://doi.org/10.1155/2022/9199703

Ali, I., Ahmed, A I A., Almogren, A., Raza, M A., Shah, S A., Khan, A., & Gani, A. (2020, January 1). Systematic Literature Review on IoT-Based Botnet Attack. Institute of Electrical and Electronics Engineers, 8(undefined), 212220-212232. https://doi.org/10.1109/access.2020.3039985

Analysis of Botnet Attack Communication Pattern Behavior on Computer Networks. (2022, June 24). Intelligent Networks and Systems Society, 15(4). https://doi.org/10.22266/ijies2022.0831.48

Ashraf, J., Keshk, M., Moustafa, N., Abdel-Basset, M., Khurshid, H., Bakhshi, A D., & Mostafa, R R. (2021, May 25). IoTBoT-IDS: A novel statistical learning-enabled botnet detection framework for protecting networks of smart cities. Elsevier BV, 72, 103041-103041. https://doi.org/10.1016/j.scs.2021.103041

Borges, J B., Medeiros, J P S., Barbosa, L P A., Ramos, H S., & Loureiro, A A F. (2023, December 1). IoT Botnet Detection Based on Anomalies of Multiscale Time Series Dynamics. IEEE Computer Society, 35(12), 12282-12294. https://doi.org/10.1109/tkde.2022.3157636

Dhirani, L L., Mukhtiar, N., Chowdhry, B S., & Newe, T. (2023, January 19). Ethical Dilemmas and Privacy Issues in Emerging Technologies: A Review. Multidisciplinary Digital Publishing Institute, 23(3), 1151-1151. https://doi.org/10.3390/s23031151

Ibrahim, W N H., Anuar, M S., Selamat, A., & Krejcar, O. (2022, January 4). BOTNET DETECTION USING INDEPENDENT COMPONENT ANALYSIS. IIUM Press, International Islamic University Malaysia, 23(1), 95-115. https://doi.org/10.31436/iiumej.v23i1.1789

Kambourakis, G., Kolias, C., & Stavrou, A. (2017, October 1). The Mirai botnet and the IoT Zombie Armies. https://doi.org/10.1109/milcom.2017.8170867

Owen, H., Zarrin, J., & Pour, S M. (2022, February 28). A Survey on Botnets, Issues, Threats, Methods, Detection and Prevention. Multidisciplinary Digital Publishing Institute, 2(1), 74-88. https://doi.org/10.3390/jcp2010006

Salim, M M., Rathore, S., & Park, J H. (2019, July 10). Distributed denial of service attacks and its defenses in IoT: a survey. Springer Science+Business Media, 76(7), 5320-5363. https://doi.org/10.1007/s11227-019-02945-z

Stahl, B C. (2021, January 1). Ethical Issues of AI. Springer International Publishing, 35-53. https://doi.org/10.1007/978-3-030-69978-9_4

Xing, Y., Shu, H., Zhao, H., Li, D., & Guo, L. (2021, April 14). Survey on Botnet Detection Techniques: Classification, Methods, and Evaluation. Hindawi Publishing Corporation, 2021, 1-24. https://doi.org/10.1155/2021/6640499

Yang, X., Guo, Z., & Mai, Z. (2022, July 1). Botnet Detection Based on Machine Learning. , 2018(undefined), 213-217. https://doi.org/10.1109/icbctis55569.2022.00056

Yerima, S Y., Alzaylaee, M K., Shajan, A., & Vinod, P. (2021, February 23). Deep Learning Techniques for Android Botnet Detection. Multidisciplinary Digital Publishing Institute, 10(4), 519-519. https://doi.org/10.3390/electronics10040519

Zhang, Z., Ning, H., Shi, F., Farha, F., Yang, X., Xu, J., Fan, Z., & Choo, K R. (2021, March 13). Artificial intelligence in cyber security: research advances, challenges, and opportunities. Springer Science+Business Media, 55(2), 1029-1053. https://doi.org/10.1007/s10462-021-09976-0

About the Author

Dr. Ramazan KOCAOĞLU, completed his PhD in Computer Engineering at Gazi University in 2017. In addition to completing many projects and providing consultancy services in the public and private sectors with the company he founded in 2018, he continues to produce products and solutions that can meet the needs of the IT sector. He is a project manager in TUBITAK 1501 and 1507 Industrial R&D Projects Support Programs. He continues his academic studies as an assistant professor at Ostim Technical University Computer Engineering Department. His research areas mainly include Next-generation Wireless Communication, Computer Network, Software Defined Network, Internet of Things (IoT), Vehicular Ad-hoc Network, Sensor Network, Mesh Network, Nano Network, Cyber Security, Open Source Systems, Intelligent Optimization Techniques.
E-mail: ramazan.kocaoglu@ostimteknik.edu.tr, **ORCID:** 0000-0002-6554-3335

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 6%.

Cyber Threat Analytics in Data Science: Intrusion Detection And Prevention Systems

Özgür TONKAL

Samsun University

To Cite This Chapter

Tonkal, O. (2024). Cyber Threat Analytics in Data Science: Intrusion Detection And Prevention Systems. In M. H. Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 97–108). ISRES Publishing.

Introduction

In recent years, the enormous increase in the volume of data in the digital world has necessitated the development of new approaches and solutions in the field of cyber security. Traditional security methods are insufficient against dynamic and constantly changing threats. At this point, data science comes into play with its powerful analyzing capabilities. By using machine learning, big data analytics and artificial intelligence techniques, it has become possible to not only detect but also predict attacks. This strong relationship between data science and cybersecurity plays a critical role in creating a more effective and proactive defence mechanism.

Data science is used to detect cyber attacks and learn the patterns of these attacks by analysing large amounts of structured and unstructured data. The data used in cyber security spans a wide range of areas such as user activities, network traffic and system logs. Analyses made on this data enable earlier detection of attacks and faster response to threats. Thus, data science techniques play a critical role in managing cyber security threats more effectively and making security solutions more flexible.

Nowadays, cyber-attacks are becoming more sophisticated and harder to detect. Traditional signature-based security systems are weak against unknown or newly derived attacks, as they can only protect against previously identified threats. Therefore, modern cyber security solutions need to be more proactive and predictive. Cyber threat analytics helps detect and prevent attacks by bringing data science into play to fulfil this need (Agbadoku, 2024).

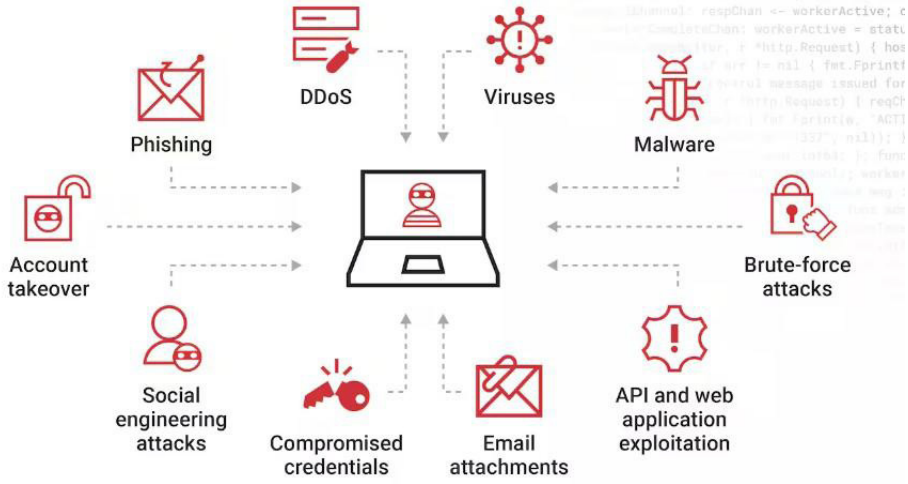
Data science offers a powerful analytical tool for identifying threat patterns and predicting potential attacks based on historical data of cyber security incidents (Babu et al., 2018). Machine learning and deep learning algorithms, in particular, can detect anomalous behaviour, go beyond known threats to recognise new types of attacks and provide real-time intervention. This enables organisations to minimise security risks and make better strategic decisions. This section will focus on how data science is used in cyber threat analytics. Firstly, data science techniques used for the detection and prevention of cyber threats will be analysed in detail. In particular, how data science approaches are integrated into critical security structures such as intrusion detection systems (IDS) and intrusion prevention systems (IPS) will be emphasised. In this context, important techniques such as machine learning, big data analytics and anomaly detection methods will be elaborated. The aim of the chapter is to understand the critical role of data science techniques in cyber security and to illustrate how these techniques are applied in practice. It also aims to provide readers with an understanding of the current state of cyber security

analytics and a glimpse of where the field may evolve in the future.

Cyber Threats and Their Categories

Cyber threats are constantly evolving and diversifying. More complex attack techniques emerge every day and new ways are discovered to bypass organizations' cybersecurity systems. Figure 1 shows cyber attack vectors. While attack vectors are increasing day by day, the main types of cyber threats are as follows: (Chakraborty et al., 2023)(Kashif et al., 2018)

Figure 1
Cyber Attack Vectors



Distributed Denial of Service Attacks (DDoS):

DDoS attacks aim to render the system unserviceable by sending excessive amounts of traffic to a specific target. This type of attack can be detected, especially with big data analytics techniques. Abnormal increases in network traffic can be analyzed by machine learning algorithms to quickly determine whether there is a DDoS attack (Tonkal et al., 2021).

Malicious Software (Malware):

Malware includes types of software that aim to cause damage by gaining unauthorized access to systems. Data science techniques, especially big data and anomaly detection methods, can detect unusual activities in file and network traffic. In this way, unknown malware can be uncovered more effectively.

Phishing:

Phishing attacks force users to share sensitive information via deceptive emails or websites. By analyzing such attacks, data science can detect recurrent patterns in the email content. Natural language processing (NLP) techniques are commonly used in the automatic classification of phishing emails (Alhogail & Alsabih, 2021).

Zero-day Attacks:

Zero-day attacks are attacks carried out before software vulnerabilities are discovered. While such attacks are extremely difficult to detect, machine learning algorithms can identify potential zero-day attacks by analyzing anomalies in network traffic.

Insider Threats:

These are threats carried out by malicious employees or those at risk of data leakage.

Behavioral analytics and anomaly detection techniques used in data science have the potential to detect these threats in advance.

In addition to the above, there are many other types of cyber threats such as SQL injection, advanced persistent threats, etc. (Figure 1). Addressing these diverse and dynamic threats requires a proactive cybersecurity approach, and data science plays a crucial role in this approach. Each type of threat offers a rich source of data that can be analyzed with data science. These data sources include a wide range of information such as system logs, user behavior, and network traffic, and can provide effective results when the right algorithms are used (Ávila et al., 2021).

Cyber Threat Analytics in Data Science

Data science plays a key role in cyber threat analytics emerging in the process of earlier detection and prevention of attacks. Machine learning, big data analytics, and other data science approaches are used to detect attacks and prevent threats proactively. Thanks to these methods, attacks that cannot be detected by traditional methods can be recognized and security strategies can be made more proactive.

Machine Learning: In cybersecurity, machine learning is one of the most critical tools used to detect threats quickly and effectively. Machine learning algorithms predict future threats and detect anomalies by learning from historical data. In particular, supervised, unsupervised, and reinforcement learning models are used to detect anomalies in network traffic and user behavior (Martínez et al., 2019).

- ***Supervised Learning***: Learning from large data sets in which attack types are labeled. For example, to identify phishing emails, models can be built that classify spam and secure emails.

- ***Unsupervised Learning***: Especially effective at detecting previously unseen threats, such as zero-day attacks. Studies unusual behavior in network traffic to detect anomalies.

- ***Reinforcement Learning:*** It is a powerful artificial intelligence method that enables systems to react faster and proactively against threats in areas such as attack detection, malware classification, and the development of dynamic defense strategies in cyber security.

Big Data Analytics: Large amounts of data need to be analyzed to detect cyber security threats. Data from sources such as network traffic, user activities, and system logs are processed and analyzed with big data tools. Big data analytics plays a critical role in understanding large-scale threats, especially DDoS attacks (Alani, 2021).

Anomaly Detection: Anomaly detection is one of the most important applications of data science techniques. Identifying deviations from normal is important for the early detection of threats. For example, a user's unusual login to the system or unusual network traffic may be a sign of a potential attack. Anomaly detection algorithms can automatically recognize these deviations and alert the system(Kaur et al., 2013).

Deep Learning: Deep learning algorithms play an important role, especially in detecting more complex attacks and analyzing data in greater depth. In particular, deep learning techniques are used effectively in advanced malware detection and the identification of anomalous network traffic patterns(Kimanzi et al., 2024).

Intrusion Detection Systems (IDS)

Intrusion Detection Systems are critical cyber security systems that monitor network and system traffic to detect potential security breaches or attacks. **Figure 2** shows the placement of IDS in the network structure. Advanced data science techniques, especially machine learning and big data analytics have become essential components of modern intrusion detection systems. IDSs are generally based on two basic approaches: signature-based and anomaly-based detection systems.

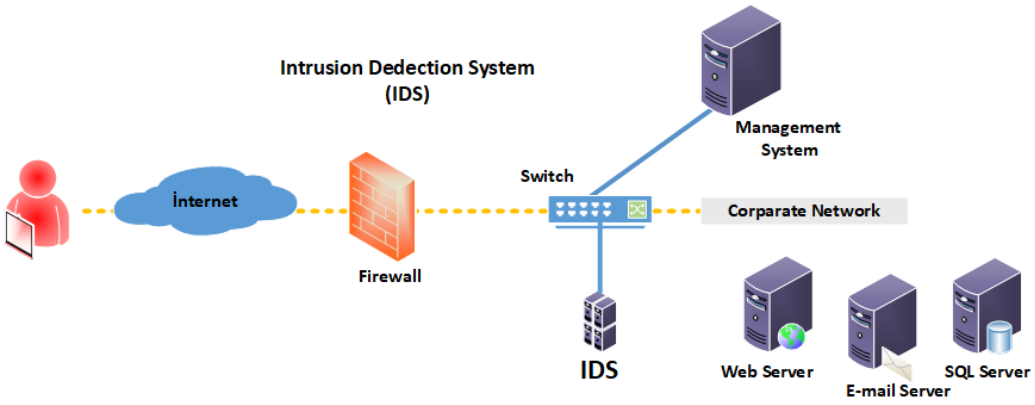
Signature-based IDS: These systems use signatures (pre-defined attack patterns) of known threats to detect threats. For example, if a particular piece of malware is infecting the system, the signature that contains the characteristics of this software will be detected by the IDS and an alert will be issued. However, signature-based systems cannot detect new threats, such as zero-day attacks, because they can only detect previously identified threats.

Anomaly-based IDS: An anomaly-based IDS learns the normal patterns of behavior of the system and issues an alert when there is a deviation from that normal behavior. Because these systems assume that any anomaly can potentially be a threat, they are more effective at detecting unknown threats.

IDS systems are generally divided into two categories: Network-based IDS (NIDS) and Host-based IDS (HIDS). **NIDS** detects threats by analyzing network traffic and **HIDS** analyses events on a specific device or system.

Figure 2

IDS Network Placement



Machine Learning Based IDSs

While classical IDS systems use specific rules and signatures to detect threats, machine learning-based IDS systems learn attack patterns by analyzing data and detecting threats accordingly. The advantages offered by machine learning-based IDSs are as follows:(Suthishni & Kumar, 2022)

- ***Detect New Attacks:*** Machine learning-based IDSs can even recognize previously unseen attacks by using the information they learn from data sets. This is the biggest advantage over signature-based systems. Unsupervised learning techniques enable these systems to detect known threats as well as zero-day attacks.
- ***Higher Accuracy Rate:*** Machine learning algorithms can analyze very large data sets and identify anomalies more accurately. By better understanding attack patterns and continuously updating the system, false positive and false negative rates are reduced.
- ***Dynamic and Adaptable Systems:*** While traditional IDSs are based on static rules, machine-learning-based systems are constantly updated with new data. This allows systems to adapt to new types of threats over time. In addition, such

systems can be continuously trained and become more effective over time.

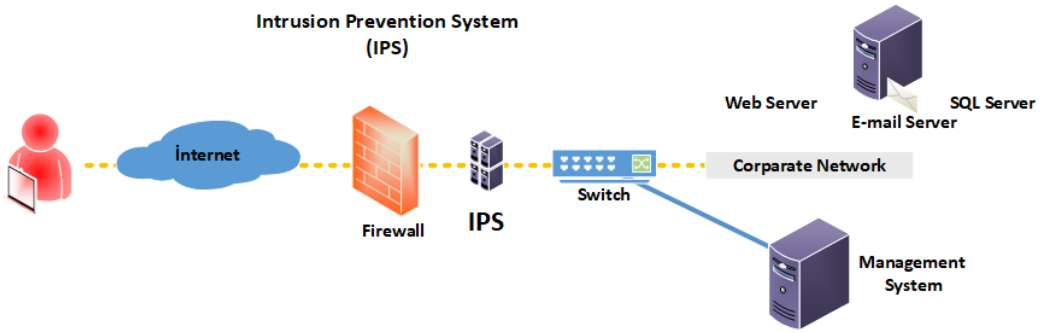
- ***Real-Time Analysis:*** IDSs powered by machine learning can quickly process large amounts of data and detect threats in real time. This provides a critical advantage, especially against fast-moving threats such as DDoS attacks.

Machine learning-based IDSs protect systems against threats using both supervised and unsupervised learning techniques. While supervised learning creates datasets for known threats, unsupervised learning attempts to find previously undetected threats through anomaly detection.

Intrusion Prevention Systems (IPS)

Intrusion Prevention Systems (IPS) not only detect threats by monitoring network traffic and system activity but also actively respond to these threats. **Figure 3** shows the placement of IPS in the network structure. When IPS identifies a potential attack, it takes steps such as stopping, blocking, or limiting harmful activity. IPSs go one step beyond IDS systems and automatically respond after detecting threats (Jayalaxmi et al., 2022). These systems can work as network-based (NIPS) and host-based (HIPS).

Figure 3
IPS Network Placement



Data Science and Proactive Defence

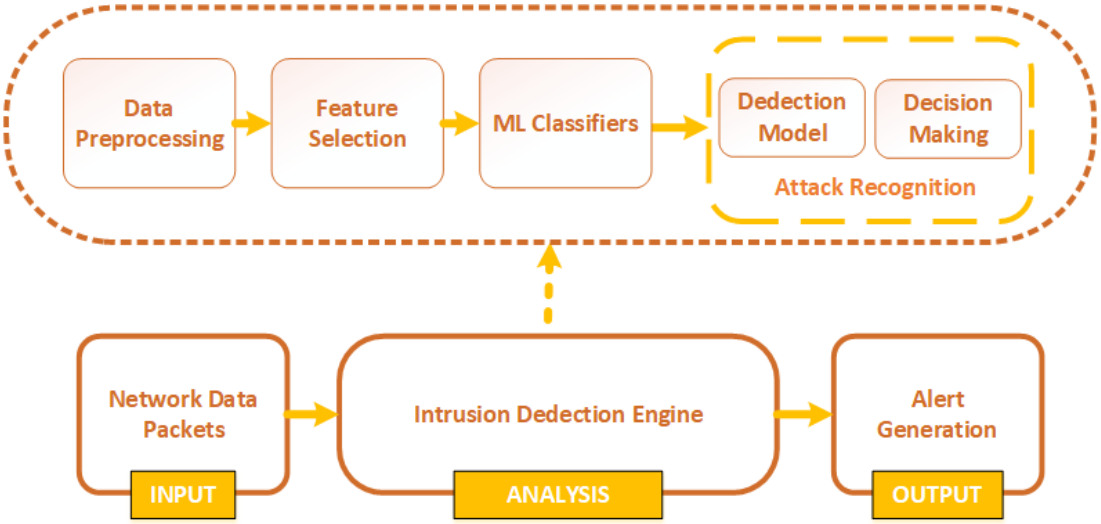
Data science tools enable attack prevention systems to be more proactive. Using machine learning, big data analytics, and anomaly detection algorithms, potential threats can be predicted before they occur. This allows systems to take faster and more accurate steps to prevent attacks. Predictive analyses and learning models, especially those used in data science, help to identify future threats in advance and close security gaps. Thus, IPS systems are not only reactive but also proactive defense mechanisms.

IPS and Deep Learning

Deep learning enables IPS systems to become more sensitive and comprehensive. Deep learning algorithms, especially working on large data sets, better analyze anomalies and are highly effective in identifying complex attack patterns. Deep learning-based IPS systems achieve significant success in detecting unknown threats and responding quickly to advanced attacks. These techniques enable systems to continuously improve themselves and create stronger defenses against new types of threats (Jayalaxmi et al., 2022).

Figure 4

Threat Analysis Process with Data Science



Explanatory Note: Figure 4 shows a model showing the application steps of these methods. The process that starts with the collection of network traffic continues with the processing of data and the training of different learning models.

The trained model can detect/block unauthorized access and inform the network administrator. The stages in the model are explained below:

Data Collection

Collecting data from different sources like network traffic, user activities, and system logs into a data lake. The main data sources are:

- ***Network Traffic:*** Packet logs and network activity provide critical data for the detection of network-based attacks.
- ***System Logs:*** User activity, transaction logs, and file access information are used to identify system threats.
- ***User Behaviour:*** Data used to detect abnormal user behavior, helping to understand insider threats.
- ***Security Incidents:*** Incident data from security incident management (SIEM) systems is an important source for threat analysis.

Data Pre-processing and Cleaning

In order to make effective analyses in threat detection, pre-processing of raw data is required. Data pre-processing stages include the following:

- ***Data Cleaning:*** Cleaning erroneous, incomplete, or duplicate data helps to achieve more accurate results.
- ***Data Transformation:*** The transformation of data into a form that can be analyzed. Especially for network traffic data, it is important to normalize the different protocols.
- ***Feature Selection:*** Selecting meaningful features from data critical to threat detection improves the performance of algorithms.

Modelling and Algorithms

Data science algorithms for threat detection and prevention work effectively on large data sets to detect attack patterns. These algorithms are used both in dealing with known threats and in detecting new and unknown attacks. Here are the three main algorithms commonly used in threat analytics and their details in cyber security applications:

Decision Trees

- Decision trees are a branching data modeling technique that reveals how to reach a conclusion based on various conditions. These algorithms create branching

points by using features in the data set and terminate the final classification at a leaf node. They are one of the most widely utilized methods for intrusion detection for the following reasons:

- ***Easy Interpretability:*** Decision trees are highly transparent in explaining how attacks are identified. By analyzing the branching points in the decision tree, security analysts can discern which decisions lead to specific conclusions.
- ***Minimal Data Preparation:*** Decision trees can operate with limited data preprocessing requirements, which is advantageous when working with large and diverse data sets.
- ***Efficiency and Speed:*** Decision trees run quickly and can effectively process extensive data sets. Moreover, they enable the instant detection of attacks by providing rapid results.
- ***Applications Areas****:* Decision trees are widely used, especially in the detection of DDoS attacks, malware classification, and phishing attacks

Support Vector Machines (SVM)

- Support Vector Machines (SVM) is an algorithm that shows high success in classification problems. SVM finds the hyperplane that provides the widest range between data points in order to classify them. Important advantages of SVM in cyber threat detection are as follows:
  - ***High Accuracy****:* SVM is very successful in large data sets containing many features thanks to its ability to work with complex data structures. It offers high accuracy, especially in problems such as network anomaly detection.
  - ***Efficiency on Small and Large Datasets****:* SVM can operate efficiently on both small and large data sets. This feature is especially advantageous in security projects with limited data sets.
  - ***Overfitting Robustness****:* SVM, when properly optimized, reduces the risk of overfitting and provides more general solutions. This is an important factor in the ever-changing nature of cybersecurity threats.
  - ***Applications Areas****:* SVM is widely used in anomaly detection, spam email detection, malware detection, and network traffic analyses.

Neural Networks

- **Neuronal networks** are a technique that works by modeling neurons in the human brain and are particularly used in deep learning algorithms. Applications of neural networks in cyber security perform better, especially on large and complex datasets:
  - ***High Performance with Deep Learning****:* Neural networks have the ability to learn more complex attack patterns with deep learning methods. It provides great success, especially in the detection of zero-day attacks. Deep learning can detect hidden patterns of attacks using multilayer neural networks.
  - ***Compatibility with Big Data****:* Neural networks work well with large and complex data sets. In systems with continuous data flow, such as cyber security, deep learning methods can constantly update themselves by learning from this data flow.
  - ***Feature Engineering****:* Neural networks, unlike other algorithms, reduce the need for feature engineering. That is, the model can learn by automatically discovering features. This eliminates the hassle of manually selecting features in big data.
  - ***Application Areas****:* Neural networks are widely used in malware detection, intrusion prevention systems, network anomaly analysis, and phishing attack detection.

All of these algorithms are used to more effectively detect and prevent cyber security threats. The advantages and weaknesses of each algorithm can be optimized for particular use cases. For example, neural networks are effective for large data sets and complex patterns, while decision trees provide more explainable and faster results. Support vector machines, on the other hand, provide accurate and robust results in complex classification tasks.

Real-World Applications and Case Studies

The success of data science techniques in detecting and preventing cyber threats has been demonstrated in many sectors and different security scenarios. Here are some examples of successful applications using these techniques (Opara et al., 2022):

Google's Security Solutions: Google uses advanced machine learning techniques to protect its users' data. In particular, machine learning-based algorithms have achieved great success in spam filtering, detecting phishing attacks, and monitoring abnormal account activity. Billions of emails are analyzed every day and advanced models are applied to protect users from fake or malicious content (Google Cloud security solutions, 2024).

Darktrace: Darktrace, an artificial intelligence and machine learning-based cybersecurity company, focuses on preventing cyberattacks by monitoring network traffic and performing anomaly detection. The company uses self-learning algorithms to detect threats from both inside and outside. Darktrace has helped protect many large companies from cyber attacks (Darktrace Security Solutions, 2024).

IBM Watson for Cyber Security: IBM is using its Watson AI system to analyze cybersecurity threats. By examining large amounts of structured and unstructured data, Watson can make fast and accurate analyses to identify threats. It is especially successful in detecting phishing attacks and ransomware. Watson analyses cyber threat intelligence data and provides recommendations to security analysts (Artificial intelligence (AI) cybersecurity, 2024).

Machine learning and anomaly detection algorithms are among the critical tools for detecting and stopping such attacks in the early stages. Table 1 lists the recent cyber security attacks and the measures taken.

Table 1
Cyber Security Attacks and The Measures Taken

Attack	The Effect	Data Science Method
WannaCry Ransomware Attack (2017) (The WannaCry ransomware attack, 2017).	It is a major ransomware attack that affected more than 200,000 systems all over the world in 2017. It targeted Microsoft Windows operating systems using a security vulnerability called EternalBlue.	**Early Detection with Machine Learning:** Machine learning-based IDS systems have been able to detect threats in the early stages of an attack by detecting unusual activities in the network. In particular, abnormal file movements and system behavior have been successfully detected by neural network algorithms. **Malware Analysis with Data Analytics:** WannaCry's propagation process was analyzed by data mining with malware analysis tools.

Target Data Breach (2013) (Pigni et al., 2017).	Target, one of the largest retail chains in the US, suffered a data breach in 2013 in which the credit card information of 40 million customers was stolen as a result of a cyber-attack.	**Anomaly Detection:** If anomaly detection systems had been able to detect abnormal data movements in the network during the attack, it may have been possible to stop the breach at an early stage. Analyses conducted after the breach revealed the importance of data anomaly monitoring techniques during the attack process. **ML to Prevent Credit Card Fraud:** In the aftermath of the attack, many financial institutions have made machine learning-based fraud detection systems more effective to prevent the misuse of stolen credit cards. These systems stopped fraudulent activity by analyzing unusual shopping activity.
SolarWinds Attack (2020) (Kruti et al., 2023).	It is a supply chain attack that took place in 2020 and affected many US government departments and large private companies. Attackers were able to infiltrate networks by placing a malicious update to SolarWinds' Orion software.	**Anomalous Behaviour Detection:** Abnormal network activity on systems using SolarWinds software was detected by data science-based anomaly detection algorithms. **Attack Pattern Recognition:** Deep learning algorithms have enabled a clearer understanding of the size of the attack by recognizing the patterns that attackers follow in networks. Especially due to the sophisticated nature of the attack, early detection of attacks has been possible with big data analyses
Colonial Pipeline Attack (2021) (Beerman et al., 2023)	It is one of the largest pipelines supplying fuel to the east coast of the USA and was attacked by ransomware in 2021. This attack, carried out by a cybercrime group called DarkSide, caused large-scale fuel disruptions in the US.	**Ransomware Detection**: Machine learning and big data analysis were used to detect ransomware infiltrating the Colonial Pipeline system. In particular, the detection of ransomware activity was based on unusual changes in system behavior during the attack. **Early Warning with Anomaly Detection**: Post-attack analysis has shown that systems can be better protected by detecting abnormal behavior on the network using big data and machine learning techniques. **Forensic Analysis After the Event**: After the attack, big data analytics was used to understand how the attack was carried out. Analyzing log data from the systems and the paths taken by the attackers revealed how the attack took place in the supply chain and which vulnerabilities were exploited.

These real-world examples illustrate how data science techniques are being used effectively to combat cyber threats. Machine learning, big data analytics, and anomaly detection techniques play a vital role in the detection and prevention of modern attacks.

Conclusion and Future Trends

This chapter provides an in-depth overview of the intersection between data science and cyber security, including cyber threat analytics, intrusion detection, and prevention systems. Firstly, the role and importance of data science in cyber security is discussed. Cyber threat analytics stands out as a critical tool to deal with modern threats; machine learning, big data, and analytical approaches offer great advantages in attack detection and prevention.

Applications such as intrusion detection systems (IDS) and intrusion prevention systems (IPS) enable the integration of data science methods. In particular, the advantages of machine learning-based systems over traditional methods, combined with anomaly detection and proactive defense mechanisms, increase the security level. Data collection, pre-processing, modeling, and algorithms are considered critical components of the threat analysis process. Real-world applications demonstrate how data science is effectively used to combat cyber threats, while case studies reveal the measures taken and solutions developed against significant attacks.

However, there are also challenges that data science faces in cyber threat analytics. Technical challenges such as big data management, data quality, and real-time processing can limit the ability of security experts to work effectively. Future directions have the potential to offer significant innovations in machine learning and artificial intelligence, autonomous security systems, and data privacy. Future studies in the field of data science and cyber security may focus on the following areas:

1. **New Machine Learning Models**: It is important to develop more advanced machine learning models to deal with new threats in cyber security. In particular, research should be conducted on how techniques such as deep learning and transfer learning can be used more effectively in anomaly detection and attack classification.

2. **Data Privacy and Ethical Approaches**: Addressing ethical issues related to user data protection and data privacy should be an important part of future work. In this area, it is recommended to develop more transparent and reliable data collection and analysis methods.

3. **Cyber Threat Intelligence and Cooperation**: Research should be carried out on increasing inter-agency cyber threat intelligence sharing and co-operation. This could enable joint defense strategies and faster response to threats.

4. **Autonomous Security Systems**: The development and implementation of autonomous systems will be an important step in the detection and prevention of cyber threats. The focus should be on increasing the ability of these systems to operate without the need for human intervention.

5. **Real-Time Analysis Methods**: Real-time data processing and analytical methods will increase the ability to respond quickly to cyber security. The development of more effective algorithms in this area can enable security analysts to be more proactive against threats.

In conclusion, this study in the field of data science and cybersecurity emphasizes the importance of cyber threat analytics and innovations in this field and prepares an important basis for future research. Cybersecurity requires the effective use of data science methods in the ever-changing threat environment and the development of innovative solutions in this field has become a great need.

References

Agbadoku, E E. (2024, January 1). The Application of Data Analytics in the Investigation of Cyberattacks: Scope and Impact. RELX Group (Netherlands). https://doi.org/10.2139/ssrn.4738358

Alani, M M. (2021, January 6). Big data in cybersecurity: a survey of applications and future trends. Springer Science+Business Media, 7(2), 85-114. https://doi.org/10.1007/s40860-020-00120-3

Alhogail, A., & Alsabih, A. (2021, July 22). Applying machine learning and natural language processing to detect phishing email. Elsevier BV, 110, 102414-102414. https://doi.org/10.1016/j.cose.2021.102414

Artificial intelligence (AI) cybersecurity. (2024, October 25). , undefined(undefined). https://www.ibm.com/ai-cybersecurit

Ávila, R L F D., Khoury, R., Khoury, R., & Petrillo, F. (2021, March 11). Use of Security Logs for Data Leak Detection: A Systematic Literature Review. Hindawi Publishing Corporation, 2021, 1-29. https://doi.org/10.1155/2021/6615899

Babu, F., Sebastian, K., & Sebastian, K. (2018, August 29). A Review on Cybersecurity Threats and Statistical Models. IOP Publishing, 396, 012029-012029. https://doi.org/10.1088/1757-899x/396/1/012029

Beerman, J T., Berent, D., Falter, Z., & Bhunia, S. (2023, May 1). A Review of Colonial Pipeline Ransomware Attack. , undefined(undefined). https://doi.org/10.1109/ccgridw59191.2023.00017

Chakraborty, A., Biswas, A., & Khan, A K. (2023, January 1). Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation. Springer Nature, 3-25. https://doi.org/10.1007/978-3-031-12419-8_1

Darktrace Security Solutions. (2024, October 25). , https://darktrace.com/platform

Google Cloud security solutions. (2024, October 25). ,https://cloud.google.com/solutions/security

Jayalaxmi, P L S., Saha, R., Kumar, G., Conti, M., & Kim, T. (2022, January 1). Machine and Deep Learning Solutions for Intrusion Detection and Prevention in IoTs: A Survey. Institute of Electrical and Electronics Engineers, 10(undefined), 121173-121192. https://doi.org/10.1109/access.2022.3220622

Kashif, M., Arshad, S., Tahir, M., Umair, M., & Waqas, P. (2018, January 1). A Systematic Review of Cyber Security and Classification of Attacks in Networks. Science and Information Organization, 9(6). https://doi.org/10.14569/ijacsa.2018.090629

Kaur, H., Singh, G., & Minhas, J. (2013, March 10). A Review of Machine Learning based Anomaly Detection Techniques. , 2(2), 185-187. https://doi.org/10.7753/ijcatr0202.1020

Kimanzi, R., Kimanga, P., Cherori, D., & Gikunda, P K. (2024, February 26). Deep Learning Algorithms Used in Intrusion Detection Systems -- A Review. Cornell University. https://doi.org/10.48550/arXiv.2402.

Kruti, A., Butt, U J., & Sulaiman, R B. (2023, January 1). A review of SolarWinds attack on Orion platform using persistent threat agents and techniques for gaining unauthorized access. Cornell University. https://doi.org/10.48550/arxiv.2308.10294

Martínez, J M., Comesaña, C I., & Nieto, P G. (2019, January 4). Review: machine learning techniques applied to cybersecurity. Springer Science+Business Media, 10(10), 2823-2836. https://doi.org/10.1007/s13042-018-00906-1

Opara, E C., Wimmer, H., & Rebman, C. (2022, July 20). Auto-ML Cyber Security Data

Analysis Using Google, Azure and IBM Cloud Platforms. , undefined(undefined). https://doi.org/10.1109/icecet55527.2022.9872782

Pigni, F., Bartosiak, M., Piccoli, G., & Ives, B. (2017, November 16). Targeting Target with a 100 million dollar data breach. SAGE Publishing, 8(1), 9-23. https://doi.org/10.1057/s41266-017-0028-0

Suthishni, D N P., & Kumar, K S. (2022, March 23). A Review on Machine Learning based Security Approaches in Intrusion Detection System. , undefined(undefined). https://doi.org/10.23919/indiacom54597.2022.9763261

The WannaCry ransomware attack. (2017, April 21). Taylor & Francis, 23(4), vii-ix. https://doi.org/10.1080/13567888.2017.1335101

Tonkal, Ö., Polat, H., Başaran, E., Cömert, Z., & Kocaoğlu, R. (2021, May 21). Machine Learning Approach Equipped with Neighbourhood Component Analysis for DDoS Attack Detection in Software-Defined Networking. Multidisciplinary Digital Publishing Institute, 10(11), 1227-1227. https://doi.org/10.3390/electronics10111227

About The Author

Özgür TONKAL received his PhD from Gazi University, Department of Computer Engineering, one of the most prestigious universities in Türkiye. He works as an Assistant Professor in the Software Engineering Department of Samsun University. His research interests include Software Defined Networks (SDN), Computer Networks, Machine Learnings, Cyber Security.

E-mail: ozgur.tonkal@samsun.edu.tr, **ORCID:** 0000-0001-7219-9053

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 7%.

Navigating the Data Science Landscape: Essential Competencies

Mehmet KOKOÇ

Trabzon University

To Cite This Chapter

Kokoç, M. (2024).Navigating the Data Science Landscape: Essential Competencies. In M. H. Calp & R. Bütüner (Eds.), *Current Studies in Data Science and Analytics* (pp. 109–123). ISRES Publishing.

Introduction

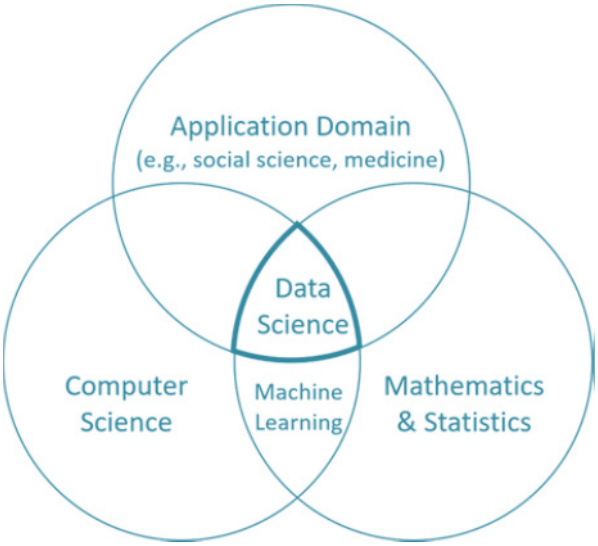
Although data science has gained prominence in recent years due to advancements in technology and business, its roots extend further back in time. Data scientists, the experts in this field, are versatile professionals who drive decision-making by extracting meaningful insights from data. To succeed, data scientists must possess a blend of technical, analytical, and soft skills. The increasing demand for data scientists has highlighted the importance of understanding the core competencies required for this profession. These skills can be categorized into technical areas such as programming, statistical and mathematical knowledge, machine learning, data management, and visualization; analytical skills like problem-solving, critical thinking, and domain expertise; and soft skills including communication, teamwork, and adaptability. However, it is important to note that these competencies can vary depending on the industry and specific role within data science. Consequently, data science training programs and curricula play a significant role in shaping these skills. Therefore, it is essential that data science education continuously evolves and adapts to industry developments. This book chapter aims to explore in detail the essential competencies data scientists must possess and contribute to the development and training processes within this field.

Data Science as a Field

Data science can be defined as an interdisciplinary field that processes structured and unstructured large data sets to derive meaningful information from them. As a field, data science integrates statistics, computer science, mathematics and domain-specific expertise to gain insights to solve complex problems as shown in the Venn diagram for data science (Hazzan & Mike, 2023). Data science is not only a process of analysis, but also includes predictions, the development of decision support systems and the optimization of processes, thus offering versatile functionality. The origins of data science can be traced back to the early 20th century and initially arose from statistical analysis. Pioneers in the field of statistics, such as Francis Galton and Karl Pearson, developed fundamental techniques such as correlation and regression (Stanton, 2001). In its modern meaning, however, data science entered a new phase with the advancement of computer technologies in the 1960s. John Tukey, in his seminal work "*The Future of Data Analysis*" published in 1962, argued that data analysis should not be limited to statistical models, but rather should be exploratory in nature (Tukey, 1962). With the advent of database management systems in the 1970s, the storage and processing of data became possible on a large scale and the concept of data mining began to take shape. With the increase in computing power in the 1980s, data analysis became more

sophisticated and enabled not only retrospective analysis but also predictive modeling. In the 1990s, the term "*data science*" officially entered the literature through the work of Peter Naur (Naur, 1992). At that time, however, data science was still considered a sub-discipline of statistics and computer science. In the 2000s, the spread of the internet and the concept of "big data" facilitated the recognition of data science as an independent field (Dhar, 2013).

Figure 1
Data Science as an Interdisciplinary Field (Hazzan & Mike, 2023)



In the 21st century, data science experienced rapid growth driven by digitization and big data technologies. Tools such as Hadoop and Spark enabled data scientists to analyze huge datasets (Zikopoulos et al., 2012). At the same time, artificial intelligence (AI) and machine learning (ML) algorithms expanded the possibilities of data science. Today, data science is not limited to understanding the past but is also a powerful tool for predicting and optimizing the future. Data science now plays a crucial role in various sectors such as healthcare, finance, marketing, education and engineering. In healthcare, for example, AI-powered models are being developed for disease diagnosis and treatment optimization, while in finance, data science is being used for fraud detection and investment strategies (Chen et al., 2021). In the future, data science is expected to focus more on automation and ethical aspects. Dealing with algorithmic bias, ensuring data privacy and promoting a more democratic use of data will be crucial to improve the sustainability of data science. Furthermore, technologies such as quantum computing could herald a new era in data science (Preskill, 2018).

Data science, which emerged at the intersection of statistics and computer science, has evolved into a scientific discipline, characterized by technological advances. During its historical development, data science has integrated various scientific fields, undergone technological transformations and achieved its current crucial importance. In the modern world, data science is transforming decision-making processes and offering innovative solutions that touch all areas of life. Although it is a concept that has gained prominence in recent years along with technological and business changes, the roots of data science lie deep in the past. Experts in the field -data scientists- extract meaningful insights from data to provide effective and timely solutions to real-world problems. So, what are the knowledge and skills that a competent data scientist needs to have? The answers to this question will be insightful for both those who want to become data scientists and educators who want to teach data science.

Competencies and Skills for Success in Data Science

Success in data science requires a blend of technical expertise (hard skills), analytical thinking and interpersonal skills (soft skills), making it a uniquely interdisciplinary field. As data becomes more complex and voluminous, the ability to gain actionable insights depends on mastering a combination of hard and soft skills. A data scientist must be proficient in programming languages, statistical methods and machine learning techniques, while also excelling in problem solving and communication. These skills not only enable the effective processing and interpretation of data, but also support collaboration across different teams and domains. In addition, ethical considerations and domain-specific knowledge are becoming increasingly important as data science impacts critical areas such as healthcare, finance and environmental studies. This section looks at the key skills that underpin data science expertise and explores their role in driving innovation and delivering impactful solutions.

The KSAO model (Knowledge, Skills, Abilities, and Other Characteristics) provides a foundational framework for understanding the competencies required for effective performance in professional roles, particularly in data science. Introduced by Boyatzis (1982) and expanded in HR and IT literature, the model identifies four key dimensions. Knowledge encompasses the theoretical and domain-specific expertise necessary for data science, such as proficiency in statistics, machine learning, and programming languages like Python and R (Binkley et al., 2012). Skills focus on the practical application of this knowledge, including data cleaning, visualization, and predictive modeling, along with soft skills like communication and teamwork (Hattingh, 2014). Abilities represent innate or developed capacities, such as problem-solving, adaptability, and critical thinking, which enable data scientists to approach complex, interdisciplinary challenges. Finally, Other Characteristics include personal traits and behavioral tendencies, such as creativity, resilience, and ethical awareness, which contribute to success in dynamic environments (Boyatzis & Kolb, 1999). This comprehensive framework allows organizations to structure training programs and performance evaluations while aligning with emerging competency models, such as KSAVE, which adds values and ethics as critical components to meet modern societal and organizational demands (Dhar, 2013; Hattingh, 2014). By integrating these dimensions, the KSAO model ensures that data scientists possess both the technical expertise and interpersonal attributes needed to navigate the complex, evolving landscape of data-driven industries.

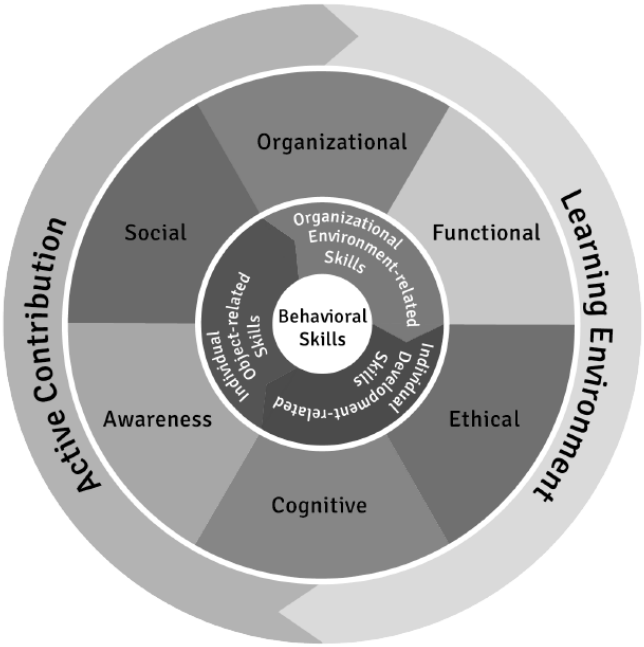
The KSAVE (Knowledge, Skills, Attitudes, Values, and Ethics) model, introduced by Binkley et al. (2012) as part of the Assessment and Teaching of 21st Century Skills (ATC21S) initiative, provides a comprehensive framework for developing skills that are critical to the modern workforce, particularly in interdisciplinary fields such as data science. KSAVE model extends traditional models such as KSAO model and incorporates values and ethics as core elements, emphasizing innovation, transparency and continuous learning alongside responsible decision making and accountability. This integration addresses the ethical and societal challenges in areas such as healthcare, finance and government where data science has a significant impact. In addition to technical competence, the model emphasizes attitudes such as curiosity, resilience and growth mindset, recognizing their importance in tackling complex problems. By prioritizing these human-centered dimensions, KSAVE model acknowledges that technical competence alone is not sufficient to meet the demands of ethical and socially responsible practices (Schleicher, 2012). The holistic approach of the KSAVE model is consistent with the interdisciplinary nature of data science and prepares professionals to use their skills responsibly and ethically. It emphasizes the importance of combining technical expertise with a strong moral framework, providing a critical foundation for the development of 21st century skills in a rapidly evolving digital landscape.

The study by Silveira et al. (2020) offers a comprehensive framework for understanding

the ***Core Skills Required in Data Science***. By employing both qualitative and quantitative methods, Silveira et al. systematically analyzed job postings to identify critical competencies for data scientists. Their proposed framework integrates six dimensions of data science, as outlined by Donoho (2017): *Data exploration and preparation, data representation and transformation, computing with data, data visualization and presentation, data modeling, and science about data science*. Each dimension encompasses both technical skills—such as programming, machine learning, and statistical analysis—and soft skills, including communication, creativity, and problem-solving. This dual emphasis underscores the multidisciplinary nature of data science, highlighting the necessity of balancing technical expertise with business acumen. Additionally, the study emphasizes the flexibility in educational requirements within the industry, reflecting a preference for practical skills over formal qualifications. By aligning academic and industry perspectives, the framework aims to guide training and recruitment efforts, thereby addressing the growing demand for skilled professionals in data science.

In the context of defining the competencies required for data scientists, a recent study (2018–2023) conducted an extensive literature review to address the lack of clarity on this topic and created ***Data Scientist Competency Framework*** (Zarefard & Marsden, 2024). The researchers developed a comprehensive framework (see Figure 2) by identifying and categorizing 130 competencies across seven critical domains: Functional, ethical, cognitive, consciousness, social, organizational and behavioral skills.

Figure 2
The Data Scientist Competency Framework (Zarefard & Marsden, 2024)



This framework, as seen in Figure 1, provides valuable insight for aligning data science education and career development with industry needs, emphasizing the interplay between technical expertise and soft skills such as communication, integration, and regulatory compliance. Functional competencies emphasize technical expertise, while ethical competencies ensure adherence to responsible practices throughout the data lifecycle. Cognitive competencies focus on intellectual skills and the use of advanced tools, while awareness competencies emphasize emotional intelligence and understanding of business contexts. Social competencies involve effective collaboration and interpersonal skills, while organizational competencies combine technical expertise with organizational goals and stakeholder relationships. Finally, behavioral competencies emphasize adaptability, entrepreneurial behavior and basic personal attributes. Together, these areas form a holistic

framework that aligns individual skills with organizational needs, encourages innovation and supports responsible and effective performance in a data-driven environment. The study highlights the importance of tailoring skills to specific roles and recognizes that it is unrealistic for one person to master all skills. This perspective is consistent with the broader understanding of data science as an interdisciplinary and collaborative field. Furthermore, the study uses an innovative methodology that combines advanced text mining and statistical analysis, setting a methodological precedent for future interdisciplinary research. By combining theoretical insights and practical applications, this study contributes to the academic discourse on data science skills while providing a robust model for assessing and improving professional performance in both academic and industrial settings.

Another study examines the data science skills, core competencies and career preferences of Nigerian college students with the aim of informing curriculum development for Library and Information Science (LIS) degree programs (Olatokun, Ayanbode, & Oladipo, 2024). According to the findings, ***success in data science requires a combination of technical and non-technical (soft) skills***. These skills are not limited to programming or data analysis skills, but also include communication, problem solving and teamwork. Research emphasizes the interdisciplinary nature of data science and points out that knowledge of programming, data management or statistics alone is not enough. To achieve meaningful results in a business context, skills such as analytical thinking, critical evaluation and domain knowledge are equally important. The study also examines the relationship between students' skills and their interest in pursuing a career in data science. It finds that the skills required for data scientists can vary depending on industry needs, highlighting the importance of LIS degree programs equipping students with both technical and non-technical skills. These findings offer valuable insights into the design of curricula for data scientists. Educational programs should include diverse learning opportunities that allow students to develop these competencies and meet industry expectations to ultimately prepare qualified data science professionals.

The previous sections have highlighted various competency frameworks and domains of expertise crucial for data scientists, including the KSAO and KSAVE models, the competency framework by Silveira et al. (2020), and the recent Data Scientist Competency Framework by Zarefard & Marsden (2024), among others. Each of these models presents unique perspectives on the essential skills for success in data science, reflecting the diversity and interdisciplinarity inherent in the field. However, upon careful examination of these competency frameworks, a common thread emerges all of these competencies can fundamentally be classified into two broad categories—technical/professional competencies and soft skills. Technical competencies encompass the core expertise in areas such as programming, machine learning, statistical analysis, data management, and other domain-specific knowledge required to perform data science tasks. Meanwhile, soft skills include interpersonal abilities such as communication, teamwork, adaptability, problem-solving, ethical considerations, and creativity, which are equally crucial for effective practice in real-world, collaborative environments.

Given the overarching categorization that these competencies naturally fall into, this book chapter aims to systematically explore data science skills under two main headings: ***Technical/Professional Competencies and Soft Skills***. By structuring the discussion in this way, I provide a comprehensive and accessible examination of the skills data scientists need, highlighting their interplay and importance in the data-driven industry. This approach allows us to critically assess how these competencies work together to support innovation, ensure ethical practice, and deliver impactful solutions in diverse fields such as healthcare, finance, and beyond.

Technical/Professional Competencies

Technical competencies are of critical importance for anyone aspiring to be a successful data scientist. These skills enable data scientists to analyze large datasets, derive

meaningful insights, and support strategic decision-making processes. This section explores the core technical skills that a data scientist must possess to excel in the field, focusing on programming, data analysis, machine learning, big data technologies, data visualization, and more.

Programming Languages

Proficiency in programming is fundamental for data scientists, as it allows them to manipulate data, build models, and create effective solutions. Python is one of the most widely used programming languages in data science, favored for its extensive libraries like Pandas, NumPy, Scikit-learn, and TensorFlow, which provide powerful tools for data manipulation and model building (Costa & Santos, 2017). SQL (Structured Query Language) is equally important, as it is essential for querying databases and retrieving data, making it a critical tool for any data scientist working with large datasets (Da Silveira et al., 2020). Additionally, R is a popular language for statistical analysis and data visualization, providing sophisticated tools for researchers and analysts to effectively explore data.

Data Analysis, Statistics, and Machine Learning

To succeed as a data scientist, it is essential to master a diverse set of competencies spanning data analysis, statistics, and machine learning. These skills encompass both theoretical knowledge and practical abilities, enabling data scientists to handle complex datasets, build models, and derive actionable insights.

Data analysis is a foundational aspect of data science, and it involves understanding, cleaning, transforming, and visualizing data to prepare it for modeling and decision-making. Exploratory Data Analysis (EDA) is one of the most important components of data analysis, as it helps data scientists understand the structure and characteristics of the data. EDA includes processes such as data cleaning, visualization, and the use of summary statistics, which are crucial for identifying patterns, detecting anomalies, and making informed decisions about the next steps in analysis (Das & Mishra, 2021; Zarbin et al., 2021). In addition, data manipulation and transformation skills are critical for converting raw data into a form suitable for analysis. This involves collecting, cleaning, organizing, and transforming data to ensure it is reliable and consistent for further use. Proficiency in data wrangling tools and techniques is essential to make raw data meaningful and actionable, supporting data-driven decision-making (Cady, 2017; Ismail & Abidin, 2016).

Statistics form the backbone of data science, providing the theoretical foundation necessary for making sense of data and developing models. A solid understanding of basic statistical concepts such as variables, sampling, correlation, and outlier handling is crucial for interpreting data and performing accurate analyses (Das & Mishra, 2021; Cady, 2017). This foundational knowledge enables data scientists to correctly apply statistical methods and derive insights that are both valid and reliable. Statistical modeling, which includes techniques such as regression analysis, hypothesis testing, and statistical inference, is another vital competency. These techniques allow data scientists to draw meaningful conclusions from data, develop predictive models, and test hypotheses, thereby enabling a deep understanding of underlying relationships within datasets (Lipovetsky, 2022; Thiruvengadam et al., 2022). For larger and more complex datasets, high-dimensional data analysis is often required. Advanced statistical techniques and data mining methods are employed to effectively analyze large datasets and extract valuable insights. Techniques suitable for high-dimensional data help data scientists handle the increasing volume, velocity, and variety of data in today's big data landscape (Cady, 2017; Donoho, 2017).

Machine learning is central to the role of a data scientist, providing the means to create predictive models and automate analytical tasks. A successful data scientist must be adept at applying basic machine learning algorithms such as k-Nearest Neighbors, Naive

Bayes, linear and logistic regression, decision trees, neural networks, and clustering. Mastery of these algorithms is necessary for building models that can classify data, predict future trends, and discover hidden patterns in datasets (Cady, 2017; Zarbin et al., 2021). Furthermore, skills in model evaluation and selection are essential for ensuring that the machine learning models developed are accurate and reliable. Techniques like cross-validation, model selection, and hyperparameter optimization play a critical role in improving the performance and robustness of models (Thiruvengadam et al., 2022). In recent years, deep learning and advanced machine learning techniques have become increasingly important, especially when dealing with large datasets and complex problems. Neural networks and deep learning methods, supported by tools like TensorFlow and PyTorch, have opened new possibilities for solving problems in image recognition, natural language processing, and other domains that involve vast amounts of data (Cady, 2017; Summa et al., 2017).

Database Management and Big Data Technologies

Competencies in database management and big data technologies are crucial for data scientists in effectively managing and analyzing large volumes of data. These skills are essential for data storage, management, accessibility, and analysis, and enable data scientists to derive meaningful insights from both structured and unstructured datasets. Database management competencies include proficiency in SQL, which is fundamental for data retrieval, manipulation, and analysis. Additionally, understanding database design helps ensure data is organized efficiently. Data scientists should also apply FAIR principles (Findable, Accessible, Interoperable, Reusable) to make data accessible and reusable across different contexts (Duke et al., 2022; Mildenberger et al., 2023). Big data technologies competencies involve using tools for big data analytics such as Hadoop and Spark to process large datasets. Proficiency in machine learning helps derive insights from big data, and familiarity with both SQL and NoSQL storage architectures ensures effective data management (Kumar et al., 2022; Nkwanyana et al., 2023). Data scientists need strong competencies in database management, big data technologies to excel in modern data environments. These competencies empower them to manage, analyze, and interpret large datasets effectively, making a meaningful impact across industries.

Data Visualization

Data scientists must possess basic knowledge of statistics and visualization, which is crucial for understanding how to best represent data (Hehman & Xie, 2021). Proficiency in programming languages like R and Python is also essential, as these are widely used for implementing visualization techniques (Reyes, 2022). Additionally, a combination of creativity and technical knowledge is needed to select the most suitable visualization types for presenting data effectively. Advanced skills include creating interactive visualizations to make data more accessible and engaging, which requires familiarity with data management technologies and tools for interactive analysis (Battle & Scheidegger, 2020). Moreover, data scientists should have a comprehensive understanding of visualization tools to create visualizations across different domains such as healthcare, IoT, and business dashboards (Shakeel et al., 2022). Educational programs and resources are essential for developing data visualization skills, preparing students for future professional roles (Kirk et al., 2021). Adopting a step-by-step approach can help data scientists learn visualization techniques effectively, particularly in programming languages like R and Python (Reyes, 2022). Visualization competencies for data scientists encompass basic statistical knowledge, proficiency in programming languages, creativity, interactive visualization skills, and data management techniques. These competencies enable data scientists to present data clearly and effectively, ensuring their analysis results are well communicated. Educational initiatives play a crucial role in the development of these skills, supporting data scientists in their professional growth.

Cloud Computing

Data scientists need a range of skills to effectively utilize cloud computing, which has become essential for processing large data sets and conducting data-intensive experiments. Core competencies include the ability to understand and utilize cloud-based services and resources required to manage analytics without relying on on-premises infrastructure (Jami & Munir, 2021). In addition, data scientists need to master the scalability and flexibility of cloud computing for processing complex data sets. This can be seen in models that use machine learning algorithms to derive insights and support decision-making processes in companies (Archana & Kamalraj, 2024). In addition, mastery of programming languages such as R and their integration with cloud APIs is crucial for performing analytics and effectively managing data on cloud platforms (Ohri, 2020). Cloud computing democratizes access to resources and enables data-driven scientific and engineering applications that were previously constrained by cost. In this context, data scientists need to understand the economic and technical dimensions of cloud services, including the pay-as-you-go model and the challenges of data transfer and storage (Simmhan et al., 2016). Data scientists need to combine technical expertise with a strategic understanding of cloud economics and the ability to integrate cloud technologies into data science workflows to maximize the potential of cloud computing in their field.

The path to becoming a successful data scientist involves mastering a diverse set of technical competencies. These include proficiency in programming languages like Python, SQL, and R, a strong foundation in statistics and machine learning, database management expertise, knowledge of big data technologies, and skills in data visualization and cloud computing. Additionally, competencies in API usage, version control, ethical considerations, and domain-specific knowledge are critical to thriving in this dynamic field (Costa & Santos, 2017; Da Silveira et al., 2020; Hattingh et al., 2019). Together, these skills empower data scientists to analyze complex datasets effectively, derive meaningful insights, and contribute significantly to strategic decision-making across industries. In the ever-evolving domain of data science, possessing these competencies is essential for driving innovation, ensuring ethical practices, and delivering impactful solutions.

Analytical Skills

Analytical and problem-solving skills are essential for data scientists, as they enable them to navigate the complexity of modern data landscapes and effectively tackle multi-layered challenges. At the heart of this expertise is ***data-driven thinking***, a mindset in which problems are approached quantitatively, and decisions are based on facts rather than intuition. This approach requires data scientists to interpret raw data, recognize patterns and derive meaningful insights that can inform strategic decisions. In addition, they must be able to break down complex problems into smaller, more manageable components that allow for systematic analysis and identification of critical variables that influence the outcome.

Another essential aspect of ***problem solving*** in data science is the ability to model real-world challenges mathematically or statistically and translate them into forms that can be analyzed computationally. Data-driven thinking further supports this by ensuring that hypotheses to specific questions are rigorously developed and tested against empirical data. This iterative process not only refines their understanding, but also guarantees that the proposed solutions are robust and based on objective analysis. Furthermore, understanding cause-and-effect relationships within datasets is crucial for distinguishing between correlation and causation, allowing data scientists to focus on eliminating causes rather than just treating symptoms.

Effective data scientists are also characterized by ***strategic thinking*** and the ability to assess risks and probabilities, both of which are enhanced by a data-driven approach. By

assessing the possible outcomes of different scenarios, they help companies anticipate challenges and seize opportunities. This includes benchmarking potential solutions, comparing their effectiveness and suggesting alternative methods if necessary. Ultimately, these capabilities are results-oriented and focus on implementing and refining solutions to ensure measurable impact. By integrating data-driven thinking with a structured, evidence-based methodology, data scientists not only solve problems effectively, but also contribute to the long-term success of the organization.

Soft Skills

The success of data scientists in the digital era increasingly depends on a combination of technical expertise and soft skills. While technical skills such as programming, data analysis, and machine learning are foundational, it is the development of soft skills that often distinguishes highly effective data scientists. Essential soft skills for data scientists include ***communication, leadership, teamwork, empathy adaptability, and problem-solving*** (Afsharian, 2024; Lu, 2024; Oropesa, 2024). These competencies are vital for effective collaboration, efficient project management, and translating complex data insights into actionable recommendations that drive business value (Desidério et al., 2024; Karneli et al., 2024).

Communication skills are particularly important because they allow data scientists to share their technical work with stakeholders who do not have a technical background. This is not just about showing numbers and charts, but rather about writing stories about specific analyses for the audiences for whom the analysis is relevant. In these disciplines, communication can mean several different things, such as a video, an infographic, an oversimplification of statistics, or the level of audience being addressed (Desidério et al, 2024). For example, when data scientists share the results of a predictive model with marketers, they are usually concerned with the application of the results to specific market segments rather than the mathematical models used. The ability to weave disparate information into a cohesive narrative is critical to ensuring that the findings are not only appreciated but also acted upon.

The ability to lead is another important skill that is more relevant today than ever. Data scientists are often required to take the lead on data-related projects, mentor junior staff and even act as advocates for data within an organization. In data science, leadership is not always just about managing people, but also about creating a roadmap of how data can be used to fulfil business needs. The focus here is on creating value through new technologies and increasing the organization's appetite for data. Data scientists who demonstrate leadership qualities describe themselves as "*agents of change*" who are actively involved in transformative processes that prescribe the inclusion of data in decision-making as well as strategic processes to improve the quality of their services (Lu, 2024).

Teamwork is also essential, as most off-the-shelf data science projects require a variety of data engineers, domain experts, stakeholders, customers and software developers. This allows them to see the business context and constraints that lead them to the data analysis relevant to that business (Coners et al., 2024). For example, when developing a predictive maintenance model for an organization, a data scientist needs to interact directly with engineers who know the equipment, operators who know about failures and provide feedback on patterns, and managers who work on the optimal allocation of resources. Such a collaborative approach ensures that the solutions developed are realistic and in line with business principles.

Empathy and flexibility are essential in a field like data science where there is a lot of change. New technologies, new tools, new ways of doing things all lead to new regulations and a data scientist should be willing to learn and unlearn as well as update their paradigm (Lu, 2024). This flexibility is also forward-looking and includes other types of innovation such as reorganization when new information has been added or the

goals of the project have changed. For example, a data scientist might use a particular data set over the course of a project, but then realize halfway through the project that external data is needed to improve the accuracy of the model. The ability to revise the analysis strategy at short notice and efficiently incorporate new sources of information can have a significant impact on the project.

Problem-solving remains at the core of data science landscape. Data scientists are, in essence, problem solvers who apply analytical skills to derive meaningful insights from data. However, problem-solving in this context often extends beyond technical challenges to include defining the right problem, understanding stakeholder needs, and ensuring that the proposed solution is feasible and valuable for the organization (Afsharian, 2024). This involves asking the right questions, breaking down complex problems into smaller, manageable parts, and systematically testing hypotheses to arrive at data-driven conclusions. For instance, when working on customer churn prediction, a data scientist must not only build a predictive model but also identify what interventions the company can realistically implement to reduce churn. Effective problem-solving in data science thus blends technical acumen with practical business considerations.

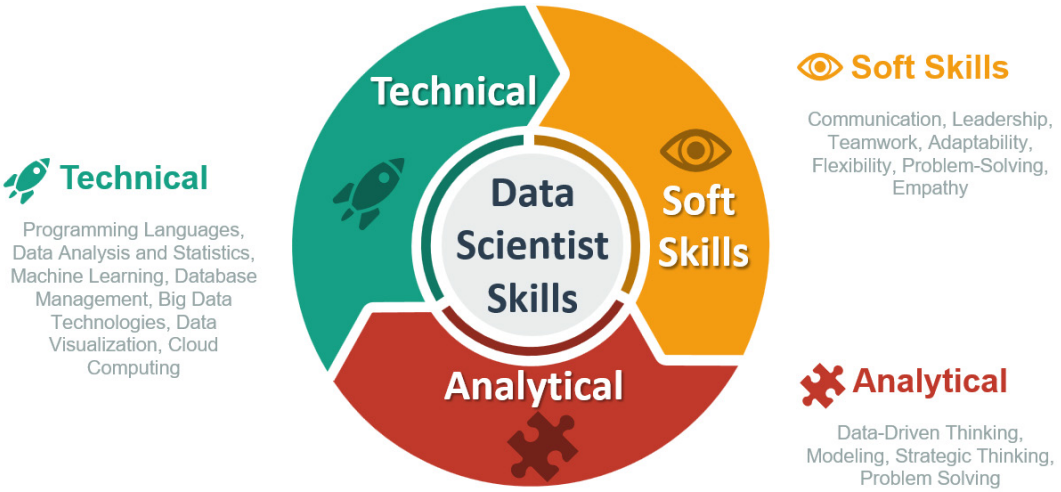
The importance of developing these soft skills cannot be understated, as they have a direct impact on project success and career advancement in the field of data science (Oropesa, 2024). Data scientists who excel at communication, leadership, adaptability, and teamwork are better equipped to lead impactful projects, influence decision-making processes, and ensure that data insights translate into tangible actions (Karneli et al., 2024). As the field of data science evolves, the role of these soft skills becomes even more critical, complementing technical capabilities to drive meaningful outcomes in the digital business landscape. Therefore, to remain competitive, data scientists must focus on enhancing both their technical and soft skills, enabling them to navigate complexities, work collaboratively, and contribute to the long-term success of their organizations (Ul Haq et al., 2024).

Ultimately, it is the integration of these soft skills — communication, leadership, teamwork, empathy and flexibility, problem solving, adaptability, and effective decision-making - that characterizes a successful data scientist. These competencies enable them to communicate their analysis persuasively, collaborate effectively and develop impactful solutions that advance their organization's strategic goals. Soft skills combined with technical know-how help turn raw data into meaningful insights that drive innovation and create significant value across industries.

Conclusion and Future Directions

Becoming a great data scientist is not an easy road as it requires a bunch of skills which includes both technical, analytical and soft skills. In this chapter, it is highlighted that data science is a seamless blend of programming, data analysis, machine learning and holistic vision to see and process complex data. In addition, it also refers to soft skills such as communication of analytical results, team cooperation effort and adaptation to the changing environment or roles in data science. The competencies of a data scientist discussed in this study are illustrated in Figure 3 for clarity and emphasis.

Figure 3
Key Skills of Data Scientist



To be considered a data scientist, one should be competent enough to understand how data can contribute to any business decision-making process. The technical skills of a data scientist primarily include programming in languages such as Python and R, machine learning modeling and statistical analysis, and working with large amounts of data. Last but not least, these individuals should also be able to visualize their results not only for themselves but also for others, regardless of technical level. In addition to the technical skills mentioned above, data scientists are also expected to have a certain level of intellect that helps them to tackle various problems effectively, e.g. the ability to break a problem down into sub-problems and solve them one by one. And just as building a team requires skill and flair, so does working in a team with members from other fields. Such successful data scientists can present their findings in writing, respond flexibly to change and make useful business recommendations based on the insights gained from the data. This ensures that regardless of the type of data available to a data scientist, whether qualitative or quantitative, they not only analyze and process data, but also implement changes in the way the organization functions and other social issues.

It is forecasted that the discipline of data science will keep pace with changes owing to advancements in technology, especially within the frameworks of artificial intelligence, cloud computing, and big data technologies. The rise of big data calls for so complex and advanced machine learning algorithms, and as the case is, quantum computing presents significant possibilities in this context. Such changes are forcing data scientists to be dynamic and progressive in their technical abilities to embrace the cutting edge of technology. In addition, it has become apparent that psychosocial issues concerning data science practice cannot be overlooked. With modern societal expectations, issues on data privacy, fairness in algorithms and ethical use of artificial intelligence are becoming prominent and data scientists must be trained in this respect. So, they also must teach future practitioners these things, hence programs will have an emphasis on ethics and values and on social claims making them civil data scientists.

Dismantling unnecessary barriers to the use of data is, and will remain, a central preoccupation of data science in the years to come. With companies increasingly accepting the need to make decisions driven by data, it becomes imperative to find ways of making data science easy for non-technical professionals, who are the majority. This serves to indicate enlarging of the audience in terms of the development of platforms and trainings designed for effective interaction with data for all concerned on a more general scale. Improving the attitude towards dealing with data and encouraging a broader category of workers to embrace data science technology, within, can, in turn, draw a greater value from data science. Helping to establish that relationship would be even more important

given the complex global issues such as climate change, health care system or socio-economic disparities require addressing. In many cases, data scientists are expected to work hand in hand with subject matter experts and leverage their collective expertise to arrive at a broad solution.

The ever-changing landscape of data science has its own rewards and challenges. To be relevant and also efficient, a data scientist is expected to possess a multitude of hard and soft skills, be open to change and evolution in technology, as well as behave in a professional and ethical manner in work. If the focus is on the promotion of life-long education, the observance of ethical principles, and the enhancement of outreach and teamwork, data science can be a very powerful force for good in many spheres and within society itself. These outlooks help data scientists build their professional paths and combine them with solving many of the demands existing in the contemporary world.

References

Afsharian, M. (2024). Data science essentials in business administration: A multidisciplinary perspective. *Decision Analytics Journal*, *11*, 100442. https://doi.org/10.1016/j.dajour.2024.100442

Archana, G., & Kamalraj, R. (2024). A Cutting-Edge Data Science Model Leveraging Cloud Computing. *International Journal of Advanced Research in Education & Technology, 11*(3), 863-869. https://doi.org/10.15680/ijarety.2024.1103006

Battle, L., & Scheidegger, C. (2020). A Structured Review of Data Management Technology for Interactive Visualization and Analysis. *IEEE Transactions on Visualization and Computer Graphics, 27*, 1128-1138. https://doi.org/10.1109/TVCG.2020.3028891

Cady, F. (2024). *The data science handbook*. John Wiley & Sons. https://doi.org/10.1002/9781119092919

Chen, J., Asch, S. M., & Asch, D. A. (2021). Machine learning and prediction in medicine — Beyond the peak of inflated expectations. *The New England Journal of Medicine, 385*(2), 190-193. https://doi.org/10.1056/NEJMp2101709

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387. https://doi.org/10.1145/362384.362685

Coners, A., Matthies, B., Vollenberg, C., & Koch, J. (2024). Data Skills for Everyone!(?)–An Approach to Assessing the Integration of Data Literacy and Data Science Competencies in Higher Education. *Journal of Statistics and Data Science Education*, 1-29. https://doi.org/10.1080/26939169.2024.2334408

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management*, *37*(6), 726-734. https://doi.org/10.1016/j.ijinfomgt.2017.07.010

da Silveira, C. C., Marcolin, C. B., da Silva, M., & Domingos, J. C. (2020). What is a Data Scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias, 8*(1), 25-39. https://doi.org/10.5585/iptec.v8i1.17263

Desidério, S. B., Lelis, M. R. L., Rodrigues, M. E., & Marques, A. B. (2024). How ready for HCI? A qualitative analysis of the practice of soft skills related to HCI by women involved in the digital girls program partners projects. *Journal on Interactive Systems*, 15(1), 504–516. https://doi.org/10.5753/jis.2024.3852

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64-

73. https://doi.org/10.1145/2500499

Donoho, D. (2017). 50 years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

Donoho, D. (2017). 50 Years of data science. *Journal of Computational and Graphical Statistics, 26*, 745 - 766. https://doi.org/10.1080/10618600.2017.1384734

Duke, R., Bhat, V., & Risko, C. (2022). Data storage architectures to accelerate chemical discovery: data accessibility for individual laboratories and the community. *Chemical Science, 13*, 13646 - 13656. https://doi.org/10.1039/d2sc05142g

Hattingh, M., Marshall, L., Holmner, M., & Naidoo, R. (2019). Data science competency in organisations: A systematic review and unified model. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019 (SAICSIT '19)*, pp. 1–8. https://doi.org/10.1145/3351108.3351110

Hazzan, O., & Mike, K. (2023). What is Data Science? In *Guide to Teaching Data Science: An Interdisciplinary Approach*, pp. 19-34. Cham: Springer International Publishing.

Hehman, E., & Xie, S. (2021). Doing Better Data Visualization. *Advances in Methods and Practices in Psychological Science, 4*(4). https://doi.org/10.1177/25152459211045334

Hersh, W., Hoyt, R., Chamberlin, S., Ancker, J., Gupta, A., & Borlawsky-Payne, T. (2023). Beyond mathematics, statistics, and programming: data science, machine learning, and artificial intelligence competencies and curricula for clinicians, informaticians, science journalists, and researchers. *Health Systems, 12*, 255 - 263. https://doi.org/10.1080/20476965.2023.2237745

Ismail, N., & Abidin, W. (2016). Data scientist skills. *IOSR Journal of Mobile Computing & Application (IOSR-JMCA), 3*(4), 52-61. https://doi.org/10.9790/0050-03045261

Jami, S. I. & Munir, S. (2021). Current Trends in Cloud Computing for Data Science Experiments. *International Journal of Cloud Applications and Computing (IJCAC), 11*(4), 80-99. https://doi.org/10.4018/IJCAC.2021100105

Karneli, O., Handayati, R., & Rijal, S. (2024). Enhancement of soft skills competence in human resources as a key success factor in the digital business era. *Journal of Contemporary Administration and Management (ADMAN)*, *2*(1), 319–324. https://doi.org/10.61100/adman.v2i1.126

Kirk, A., Santos, B., & Alford, G. (2021). A recipe of capabilities for pursuing expertise in data visualization: A practitioner's perspective. *IEEE Computer Graphics and Applications, 41*, 58-62. https://doi.org/10.1109/MCG.2020.3034737

Kumar, A., Deutsch, A., Gupta, A., Papakonstantinou, Y., Salimi, B., & Vianu, V. (2022). Database Education at UC San Diego. *ACM SIGMOD Record, 51*, 43 - 46. https://doi.org/10.1145/3572751.3572763

Lipovetsky, S. (2022). Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. *Technometrics, 64*, 145 - 148. https://doi.org/10.1080/00401706.2021.2020521

Lu, J. (2024). Data Scientist Knowledge and Skills Evaluation Towards a Data-Driven Research Methodology. In *Proceedings of the 23rd European Conference on Research Methodology for Business and Management Studies*, pp.136-144. https://doi.org/10.34190/ecrm.23.1.2321

Mildenberger, T., Braschler, M., Ruckstuhl, A., Vorburger, R., & Stockinger, K. (2023). The Role of Data Scientists in Modern Enterprises - Experience from

Data Science Education. *ACM SIGMOD Record, 52*, 48 - 52. https://doi.org/10.1145/3615952.3615966

Naur, P. (1992). *Computing: A human activity*. ACM Press/Addison-Wesley Publishing Co. https://link.springer.com/book/10.1007/978-1-4612-3096-2

Nkwanyana, A., Mathews, V., Zachary, I., & Bhayani, V. (2023). Skills and competencies in health data analytics for health professionals: A scoping review protocol. *BMJ Open, 13*. https://doi.org/10.1136/bmjopen-2022-070596

Ohri, A. (2020). *R for cloud computing: An approach for data scientists*. Springer.

Olatokun, W. M., Ayanbode, O. F., & Oladipo, S. O. (2024). Data science career preference of Nigeria University students. *Education and Information Technologies*, 1-25. https://doi.org/10.1007/s10639-024-12897-4

Oropesa, C. M. (2024). The essential soft skills of project managers and project success. Scientific *Journal of Applied Social and Clinical Science, 4*(20), 1-8. https://orcid.org/0000-0001-9168-998X

Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum, 2*(1), 79. https://doi.org/10.22331/q-2018-08-06-79

Reyes, J. M. M. (2022). *Data Visualization for Social and Policy Research: A Step-by-step Approach Using R and Python*. Cambridge University Press.

Shakeel, H., et al. (2022). A Comprehensive State-of-the-Art Survey on Data Visualization Tools. IEEE Access, 10, 96581-96601. https://doi.org/10.1109/ACCESS.2022.3205115

Simmhan, Y., Ramakrishnan, L., Antoniu, G., & Goble, C. (2016). Cloud computing for data-driven science and engineering. *Concurrency and Computation: Practice and Experience, 28*(4), 947-949. https://doi.org/10.1002/cpe.3668

Summa, M.G., Bottou, L., Goldfarb, B., Murtagh, F., Pardoux, C., & Touati, M. (Eds.). (2017). *Statistical Learning and Data Science*. Chapman and Hall/CRC. https://doi.org/10.1201/b11429

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, *9*(3). http://dx.doi.org/10.1080/10691898.2001.11910537

Thiruvengadam, K., Watson, B., Chinnaiyan, P., & Krishnan, R. (2022). A review of statistical modelling and machine learning in analytical problems. *International Journal of Applied Engineering Research, 17*(5), 506-510. https://doi.org/10.37622/ijaer/17.5.2022.506-510

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. https://doi.org/10.1214/aoms/1177704711

Ul Haq, M. U., Frazzetto, P., Sperduti, A., & Da San Martino, G. (2024, April). Improving soft skill extraction via data augmentation and embedding manipulation. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pp. 987-996.

Zarbin, M., Lee, A., Keane, P., & Chiang, M. (2021). Data science in translational vision science and technology. *Translational Vision Science & Technology, 10*. https://doi.org/10.1167/tvst.10.8.20

Zarefard, M., Marsden, N. (2024). The Essential Competencies of Data Scientists: A Framework for Hiring and Training. In: Mori, H., Asahi, Y. (eds) *Human Interface and the Management of Information. HCII 2024. Lecture Notes in Computer Science*, vol 14691. Springer, Cham. https://doi.org/10.1007/978-3-031-60125-5_27

Zikopoulos, P. C., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill.

About The Authors

M. KOKOÇ received a Ph.D. degree from the department of Computer Education and Instructional Technologies at Hacettepe University, a highly esteemed institution in Türkiye. He currently holds the position of Associate Professor in the Department of Management Information Systems at the School of Applied Sciences at Trabzon University. His research interests include Management Information Systems, E-Learning, Learning Technologies, Learning Analytics, Cognitive Profiling, Video Lectures, Human-Computer Interaction, and Cognition and Media.
E-mail: kokoc@trabzon.edu.tr **ORCID:** 0000-0002-1347-8033

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 4%.

Analysis of IoT Security Datasets

Erdal ÖZDOĞAN
Uludağ University

Onur CERAN
Gazi University

To Cite This Chapter

Özdoğan, E., & Ceran, O. (2024). Analysis of IoT Security Datasets. In M. H. Calp & R. Butuner (Eds.), *Current Studies in Data Science and Analytics* (pp. 124–143). ISRES Publishing.

Introduction

The Internet of Things (IoT) is a revolutionary technology characterized by vast data volumes across various domains. IoT systems continuously generate data through different sensors and devices, creating large data sets. This situation presents various challenges in data analysis and management. These large data volumes' processing, storage, and analysis are crucial for enhancing efficiency and obtaining meaningful insights. However, it is essential to consider that alongside the opportunities presented by IoT, various data security concerns also arise (Maiwada et al., 2024).

The security of IoT systems fundamentally relies on the protection of the data and devices present within these systems (Alrayes et al., 2024). Various data security concerns can pose threats to user privacy, data integrity, and the system itself. Among these concerns, attacks specifically designed for IoT networks hold considerable significance. Such attacks can target vulnerabilities in devices, leading to data breaches, unauthorized access, and service disruptions. Therefore, preventing these threats and mitigating their impacts plays a vital role in ensuring the security of IoT systems.

Effective solutions are required for the prevention and detection of attacks. In this context, Intrusion Detection Systems (IDS) can be utilized to secure IoT environments (Aggarwal & Sharma, 2015). IDS systems monitor network traffic, possessing the capability to detect unusual behaviors and take measures against these behaviors. The implementation of such a protection mechanism in IoT systems is crucial for enhancing the security of devices and data(Liang et al., 2024). IDSs offer a highly effective method for identifying different types of attacks and developing strategies against them.

Advanced IDS systems, when supported by machine learning-based approaches, possess strong protective potential (Bansal & Singhrova, 2023), (Saadouni et al., 2024). Machine learning algorithms can quickly identify anomalous behaviors due to their ability to analyze large data sets. The use of IoT data sets enhances the effectiveness of machine learning models. When enriched with information obtained from various sources, such as inter-device communication and user interactions, it becomes possible to predict and prevent attacks in advance.

Machine learning-based IDS systems both increase efficiency and adopt a more proactive approach against potential security threats. The utilization of such systems is not limited to the detection of security threats; they also contribute to mitigating the impacts of attacks and continuously improving the security of the system. Furthermore, the applications of these systems enhance the secure management of data, thereby raising the overall

security level of IoT systems. In conclusion, IoT systems are characterized by large data volumes and various data security concerns, presenting both opportunities and threats. Attacks unique to IoT threaten the security of these systems, while IDS systems provide a critical solution for protection against these threats. The strong protective potential of machine learning-based IDS can be further strengthened using IoT data sets. Thus, the security of IoT systems can be enhanced, and data management can be carried out more effectively.

The objective of the Research and Hypothesis

This chapter aims to explore the datasets commonly referenced in IDS systems used for IoT security in detail. Given the complexities of the networks formed by numerous interconnected devices, effective data analysis and attack detection are critical for ensuring their security. However, the effectiveness of various datasets cannot be fully assessed without conducting comparative analyses. In this context, the key concepts highlighted in this chapter will guide the examination of the datasets, aiming to identify the most suitable datasets for IoT IDS systems and contribute to research in this field.

Hypothesis 1: Implementing machine learning-based IDS will improve the security of IoT systems by enhancing the detection and prevention of various cyber threats.

Hypothesis 2: A comprehensive introduction of the datasets used in the field of IoT security will facilitate cybersecurity researchers in making more informed decisions regarding attack detection and prevention, thereby increasing the effectiveness of their research activities.

Contribution

This study will comprehensively examine and analyze commonly referenced data sets used in IDS systems for IoT security. IoT systems are characterized by complex network structures formed by the convergence of numerous devices. Effective data analysis and attack detection are critically important for ensuring the security of these structures. However, the effectiveness of various data sets cannot be fully evaluated unless a comparative analysis is conducted. In this context, the chapter will explore data sets based on key concepts that stand out. This aims to identify the most suitable data sets for IoT IDS systems and contribute to research in this field.
In this scope, the contributions of this book chapter can be summarized as follows:

- Multifaceted Analysis of Data Sets: Critical data sets for IoT security are analyzed through various dimensions and labeling types, determining in which scenarios these data sets can be utilized more effectively.
- Identification of Most Effective Features: The pre-processing needs of different data sets have been examined, and the most effective features that positively contribute to model performance have been revealed. This facilitates the development of more optimized data processing methods for IoT security.
- Diversity of Attacks: The diversity of attacks presents in the data sets used for IoT security has been investigated. This analysis will aid security researchers conducting research on specific attack types in making more informed data set preferences.

The remainder of the book chapter is organized as follows: Chapter 2 addresses current studies conducted in the field. Chapter 3 examines the role of data sets in IoT security. Chapter 4 investigates commonly used data sets and compares them within the framework of key concepts. The evaluation and conclusion section summarizes the insights obtained from the study.

Related Work

In recent years, numerous academic studies and research related to IoT-IDS, particularly

those based on machine learning, have been conducted and continue to be undertaken. The increase in the use of IoT, along with the rapid updates of machine learning models, has necessitated a high volume of academic work in this field.

In a study conducted in 2024 (Ozdogan, 2024), the effects of data preprocessing and feature selection on the simplification of machine learning algorithm selection in IoT IDS systems were analyzed in detail. In the study, which utilized multiple datasets, the performance of Machine Learning algorithms was compared from various aspects, including the preprocessing process of the datasets, the process of being balanced, and whether they were generated in a synthetic or natural environment. In the work of, the authors presented a comparative analysis of the IoT datasets used for model training, identifying key features that assist in evaluating their suitability in specific scenarios. Hota and Shrivas utilized different feature selection techniques on the NSL-KDD dataset to contribute to making Intrusion Detection Systems more efficient and effective (Hota & Shrivas, 2014). In another study conducted on the same dataset, Vibhute et al. used the NSL-KDD dataset to develop a network intrusion detection system. The study proposed and implemented a community learning-supported random forest algorithm to select the most suitable features (Vibhute et al., 2024). In his study using the NSL-KDD and UNSW-NB15 datasets, Türk conducted binary and multi-class intrusion detection experiments and achieved high-performance results (Türk, 2023). In another study using the NSL-KDD dataset, a model was created by reducing the number of features to ten using a feature selection method. To enhance prediction performance, imbalanced data was corrected using the SMOTE method (Thana-Aksaneekorn et al., 2024). In another study, the authors proposed a new IDS based on Artificial Neural Networks using the NSL-KDD dataset. The developed model was compared in terms of performance with several classifiers and achieved high success (Alrayes et al., 2024). In another study (Zoghi & Serpen, 2024), the impact of class imbalance and data overlap issues in the UNSW-NB15 dataset on the performance of data-driven models were examined. To improve classifier performance, a scalable overlap visualization method capable of detecting these issues was proposed, and its accuracy was tested with various classifiers.

A recent study using the Bot-IoT dataset proposes a new approach that combines deep learning and three-tiered algorithms to quickly and accurately detect attacks in IoT networks. Evaluations have shown that this method significantly improves detection performance compared to existing methods (Alosaimi & Almutairi, 2023).

In a study utilizing the IoTID20 and BoT-IoT datasets, a hybrid method combining PCA and the Bat Optimization Algorithm (BAT) has been proposed for dimensionality reduction in the datasets (Karamollaoğlu et al., 2024). In the study that achieved high performance, detailed analyses were conducted to determine the effects of dimensionality reduction and data balancing models on classification performance.

In another study that evaluated the performance comparisons of machine learning models, a comparative analysis of various algorithms on the Bot-IoT dataset was presented (Mishra et al., 2023). In another comprehensive study, the aim was to categorize and analyze existing datasets to create future datasets, thereby enhancing the effectiveness of intrusion detection systems and accurately reflecting network threats (Zafar Iqbal Khan et al., 2024).

Upon reviewing the studies conducted in recent years on various datasets, it can be seen that machine learning and deep learning techniques have been applied to detect cyber-attacks in IoT networks. The studies particularly focus on topics such as data preprocessing, feature selection, class imbalance, and dimensionality reduction, aiming to improve model performance. Additionally, comparative analyses of various algorithms are conducted to identify the most effective methods. Therefore, examining IoT security datasets from a data analysis perspective will provide an opportunity to test the effectiveness of new model approaches and algorithms. Furthermore, analyzing different datasets is important for understanding different types of attacks and anomalies,

leading to more comprehensive and generalizable results.

The Role of Datasets in IoT Security

In the context of IoT security, datasets play a critical role in developing effective IDS and enhancing overall system resilience against various cyber threats. The vast array of devices connected to the Internet of Things generates a large amount of data that can be leveraged for security analysis and threat detection. Data sets provide the raw data necessary to understand events and activities occurring within IoT networks. These data sets are used for training machine learning and deep learning models, directly influencing the effectiveness of these models. By utilizing diverse and high-quality data sets, researchers can develop more accurate and reliable models that enhance the security and functionality of IoT systems. The quality and representativeness of the data sets play a crucial role in enabling these models to generalize well to real-world scenarios, ultimately improving the detection and response to various threats and anomalies in IoT environments (Kaur et al., 2023).

A well-structured dataset that encompasses diverse attack scenarios allows researchers and practitioners to develop and evaluate models that can accurately identify malicious activities. For instance, datasets containing labeled instances of both benign and malicious traffic enable supervised learning techniques, which are essential for developing robust detection algorithms.

The effectiveness of an IDS is heavily reliant on the authenticity of the dataset used for training. Datasets that mimic real-world conditions, including variations in traffic patterns, device types, and attack methodologies, enhance the model's ability to generalize and perform effectively in live environments. The current methods used for labeling existing IoT datasets are based on generating synthetic network data, which overlooks the essential aspects needed to differentiate between normal and malicious behaviors (Guerra et al., 2022). Thus, the inclusion of realistic scenarios in datasets is paramount for preparing systems to confront actual threats.

Many datasets face the challenge of class imbalance, where certain attack types are underrepresented compared to benign instances (Qing et al., 2024). This imbalance can lead to biased models that fail to detect less frequent but critical threats. Effective preprocessing techniques, such as oversampling or under-sampling, can help mitigate these issues, ensuring that models are trained on a balanced representation of all classes.

Datasets allow for comparative studies that help identify the strengths and weaknesses of various IDS approaches. By evaluating multiple algorithms against the same dataset, researchers can determine which methods yield the highest accuracy and efficiency in detecting specific types of attacks (Ozdogan, 2024). This comparative analysis contributes to the continuous improvement of detection techniques and enhances the overall security posture of IoT systems.

The establishment of standardized datasets fosters a common ground for research in IoT security. These datasets serve as benchmarks for assessing the performance of IDS algorithms, facilitating a clearer understanding of advancements in the field (Neto et al., 2023). By utilizing widely recognized datasets, researchers can share their findings more effectively, driving innovation and collaboration in the cybersecurity community.

In summary, datasets are indispensable in the realm of IoT security, providing the necessary foundation for developing, testing, and refining intrusion detection methodologies. The careful selection and use of datasets impact the effectiveness of security measures, underscoring their pivotal role in safeguarding IoT environments.

Data Analysis of Datasets Used in IoT Security

The datasets used in IoT security have evolved alongside technological advancements.

The first generation of datasets typically featured simple structures with a limited number of devices and types of attacks. Over time, these datasets have become more complex, creating extensive data pools that encompass a greater variety of device types, various attack vectors, and different use case scenarios.

The primary datasets used for research and development in the IoT field generally focus on network traffic analysis, anomaly detection, classification of attack types, and the security of IoT devices. These datasets vary according to different network environments, types of attacks, and feature engineering requirements.

NSL-KDD

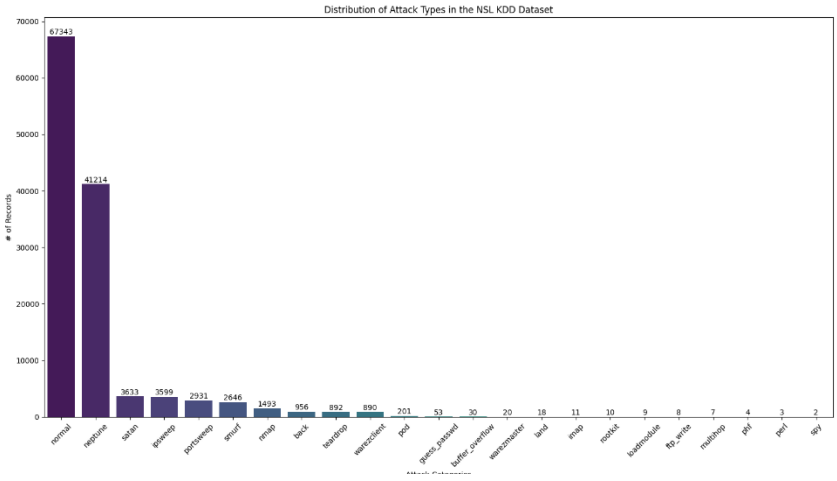
The NSL-KDD dataset is a widely used benchmark for network-based intrusion detection systems (Alosaimi & Almutairi, 2023). Developed in 2009, it addresses several issues found in the original KDD'99 dataset. One of its key improvements is the removal of duplicate and redundant records in both the training and testing subsets. This helps prevent classifiers from being biased towards more frequently occurring examples. Like the original KDD'99 dataset, the NSL-KDD dataset was created in a controlled laboratory environment. It includes records generated from a simulated environment that features various types of attacks as well as normal network traffic.

The NSL-KDD dataset consists of 41 features, each representing a different aspect of network behavior. It is divided into two main subsets Train and Test. The training dataset contains 125,973 and the testing dataset includes 22,544 records.

This dataset features a diverse range of attack types. Figure 1 illustrates the distribution of traffic types in the training dataset.

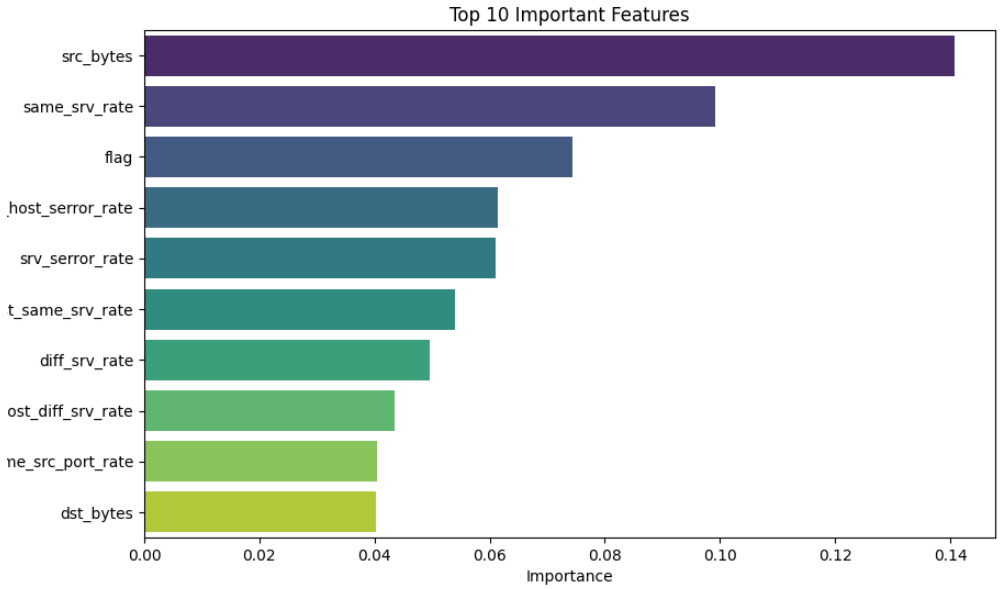
Figure 1

Distribution of Traffic Types in the NSL KDD Training Dataset



In the traffic distribution table, which is imbalanced due to the overwhelming presence of normal traffic, it is evident that the most common attack is the Neptune attack. The most effective features for attack classification are illustrated in Figure 2.

Figure 2
The Top 10 Important Features for Attack Classification of NSL KDD Dataset.



In the dataset, the most effective feature for classification is src_bytes, which indicates the amount of source bytes.

The NSL-KDD dataset provides an effective benchmark for comparing various intrusion detection methods.

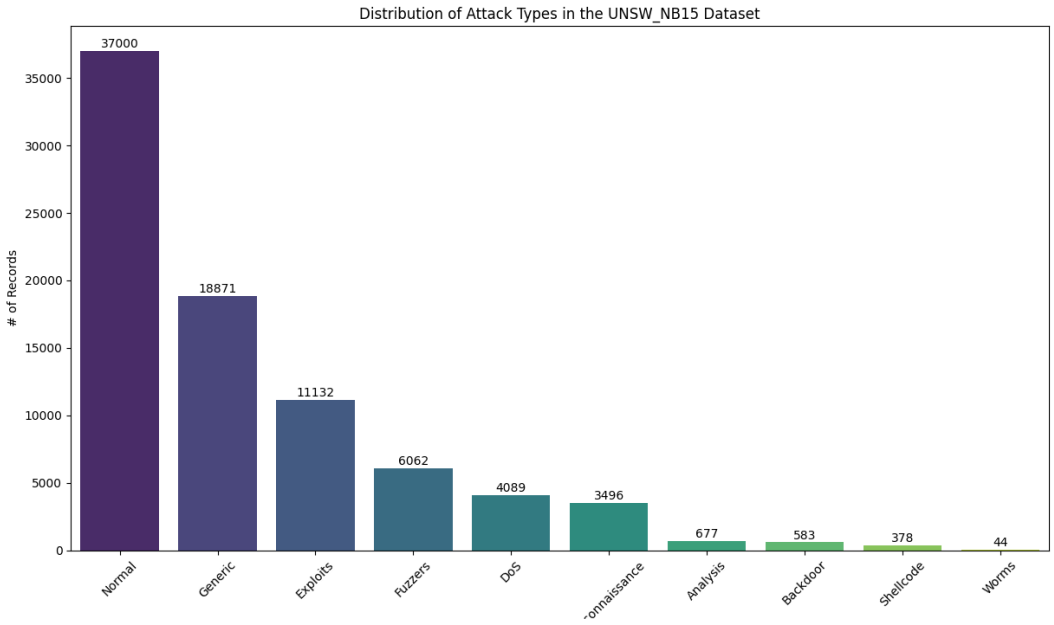
UNSW-NB15

The UNSW-NB15 dataset serves as a critical resource for the evaluation of network intrusion detection systems. This dataset was generated using the IXIA PerfectStorm tool and encompasses a comprehensive array of modern normal activities alongside various malicious attack behaviors (Moustafa & Slay, 2015, 2016).

Comprising a total of 2,540,044 records, the dataset is structured to facilitate division into training and testing subsets, with additional subdivisions available for specific research applications. The UNSW-NB15 dataset was meticulously produced in a controlled laboratory environment, specifically collected from a simulated network setup at the Cyber Range Lab located in Canberra, Australia. The network traffic within this dataset is intentionally designed to reflect contemporary attack vectors. In this unbalanced dataset, traffic is categorized as either normal or attack, with the latter further delineated into multiple attack types.

An analysis of the training subset, which consists of 82,332 records, reveals the distribution of attack types, as illustrated in Figure 3.

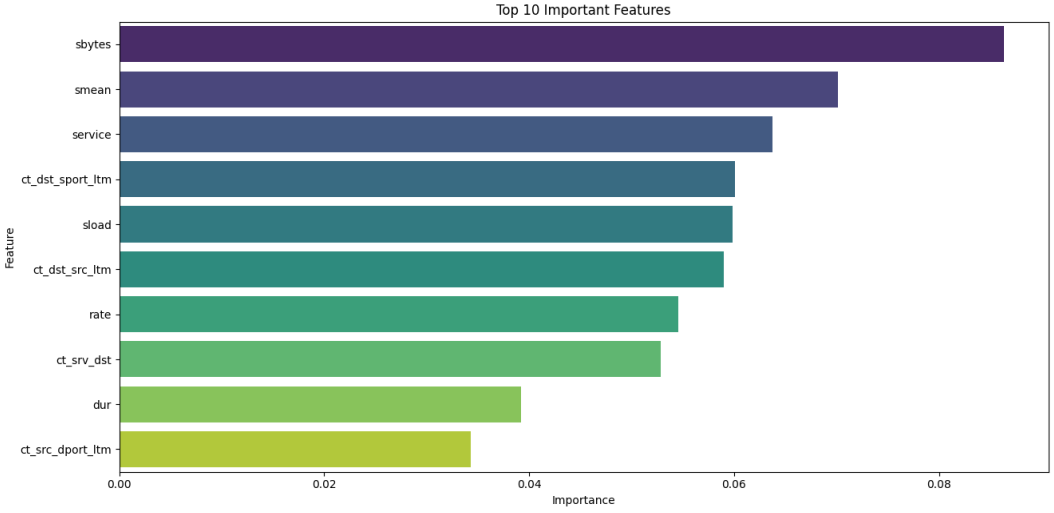
Figure 3
Distribution of Traffic Types in the UNSW-NB15 Dataset



As illustrated in Figure 3, the UNSW-NB15 dataset exhibits significant class imbalance. The distribution of different attack types and normal traffic records is not uniform, leading to discrepancies in the representation of classes within the dataset.

The classification of the traffic data using the Random Forest algorithm identifies the ten most important features contributing to this process, as depicted in Figure 4.

Figure 4
The Best 10 Features for Classification of UNSW-NB15 Dataset



The Random Forest model generates an importance score for each feature based on metrics such as total Gini decrease or entropy gain. This score indicates the model's effectiveness in utilizing that feature for classification tasks. Accordingly, the most important feature in the classification of attacks is the sbytes feature, which represents the number of source bytes. The remaining values and their impact ratios are presented in Figure 4.

The UNSW-NB15 dataset encompasses more contemporary and sophisticated attack techniques compared to earlier datasets such as KDD99 and NSL-KDD. Consequently,

it is frequently employed in cybersecurity research, particularly in domains such as IoT, network security, and intrusion detection systems.

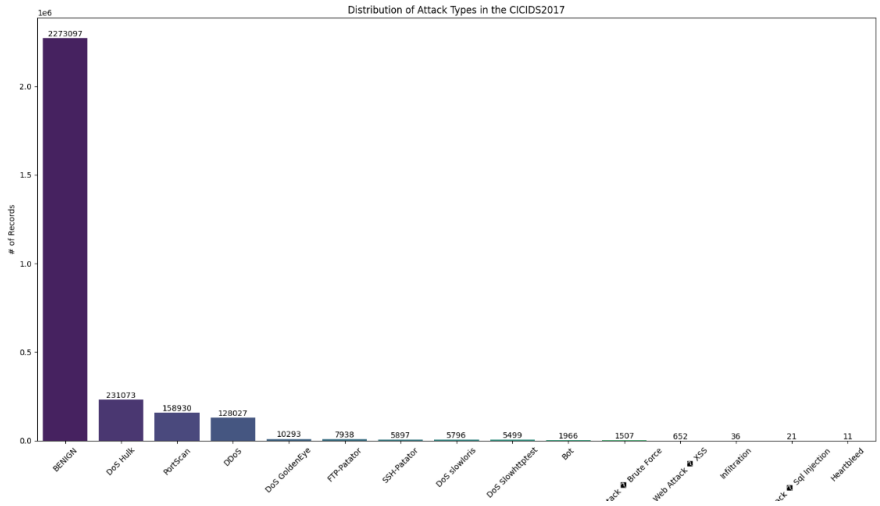
CICIDS2017

The CICIDS2017 dataset was developed by the Canadian Institute for Cybersecurity to facilitate network security research (Sharafaldin et al., 2018). Although not specifically created for IoT devices, it can be adapted for IoT IDS studies. This dataset is commonly utilized in general IDS and network security research. The dataset comprises a total of 2,830,540 records, which were generated in a simulated network environment over seven days, encompassing various attack types and normal traffic patterns that reflect real-world conditions.

The CICIDS2017 dataset includes 83 features designed to characterize each traffic instance. These features provide detailed insights into network traffic, incorporating various parameters such as protocol information, connection duration, and packet size.

An analysis of the dataset reveals the distribution of attack and benign (non-attack) traffic across different classes, as illustrated in Figure 5.

Figure 5
Distribution of Attack Types in The CICIDS-2017 Dataset



As illustrated in Figure 5, the CICIDS2017 dataset exhibits a significant imbalance, with most traffic records classified as benign (non-attack) traffic. When focusing solely on attack traffic, it is evident that the most prevalent attack categories are Denial of Service (DoS) Hulk, Port Scan, and Distributed DoS, while the Heartbleed attack demonstrates the lowest frequency. This imbalance presents challenges for the model in learning to identify such low-frequency attacks effectively.

In scenarios where there is a substantial class imbalance, models tend to favor learning the majority class (e.g., “Benign”). This tendency can hinder the accurate classification of the minority class. Additionally, classes with a low number of samples are at risk of overfitting, as the model may memorize these few instances instead of generalizing them. Consequently, while the model may perform well on the limited examples from these minority classes, it could fail when confronted with new and previously unseen instances.

TON_IoT

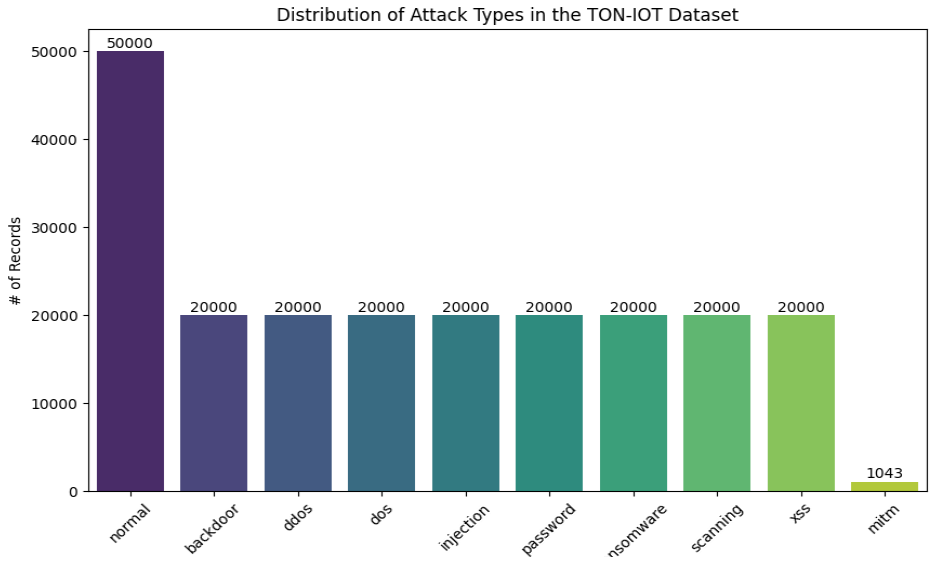
The ToN-IoT dataset, developed by UNSW Canberra, represents a next-generation resource utilized for evaluating security applications within Industry 4.0, the IoT, and Industrial IoT (IIoT) networks. This dataset serves as a benchmark for testing the accuracy and effectiveness of various cybersecurity applications, including intrusion detection systems, threat intelligence, malware detection, fraud detection, and digital forensics (Alsaedi et al., 2020; Booij et al., 2022; Moustafa, 2021).

The ToN-IoT dataset comprises a total of 12 distinct features for each record. These features encapsulate a range of parameters related to network traffic, facilitating the analysis of communication dynamics among devices. With approximately 2.8 million records, the dataset encompasses both normal traffic patterns and a variety of attack scenarios. The ToN-IoT dataset was created in a laboratory setting, where data was collected in a simulated environment that closely mirrors real-world conditions, utilizing various IoT devices (Ashraf et al., 2021; Moustafa, 2019; Moustafa, Ahmed, et al., 2020; Moustafa, Keshky, et al., 2020).

In this study, analyses were conducted using the Train Test Network dataset presented by the authors. This dataset consists of 211,043 records and encompasses 42 features. Notably, it includes not only normal traffic but also nine categories of attacks. The distribution of traffic classes is illustrated in Figure 6.

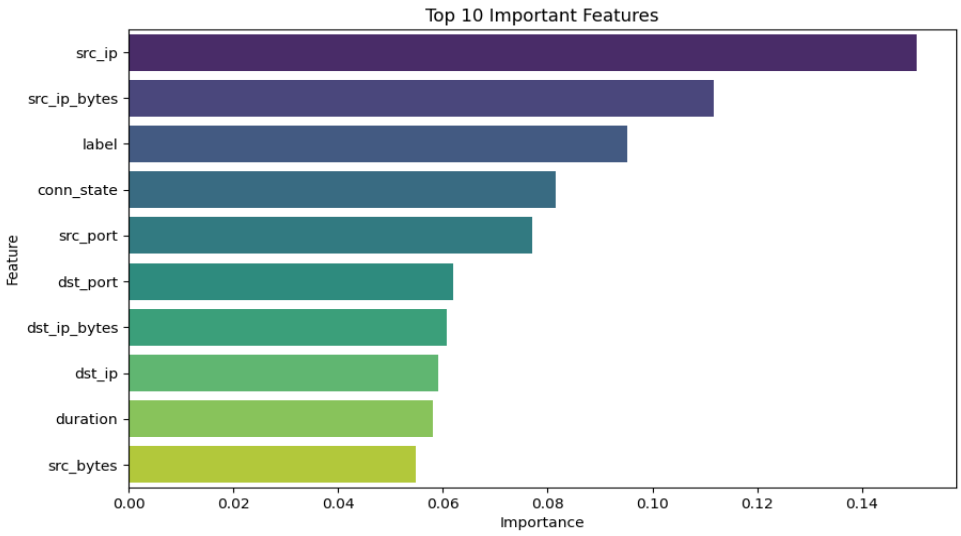
Figure 6

The Relatively Balanced Distribution of Attack Traffic in the TON-IoT Dataset.



In the dataset designed for training and testing, it has been observed that the distribution of normal traffic and all attack types, except for Man-in-the-Middle attacks, is balanced. The ten most effective features for classification purposes are illustrated in Figure 7.

Figure 7
The Top 10 Most Effective Features for Classification in the ToN-IoT Dataset.



According to the analysis, the most decisive features are primarily the source IP address, followed by features indicating the number of bytes associated with the source IP address. The TON-IoT datasets comprise telemetry data collected from IoT and IIoT sensors, as well as data obtained from Windows and Ubuntu operating systems, in addition to network traffic data. These datasets are collected from realistic and large-scale networks and encompass a variety of normal and cyber-attack events.

BoT-IoT

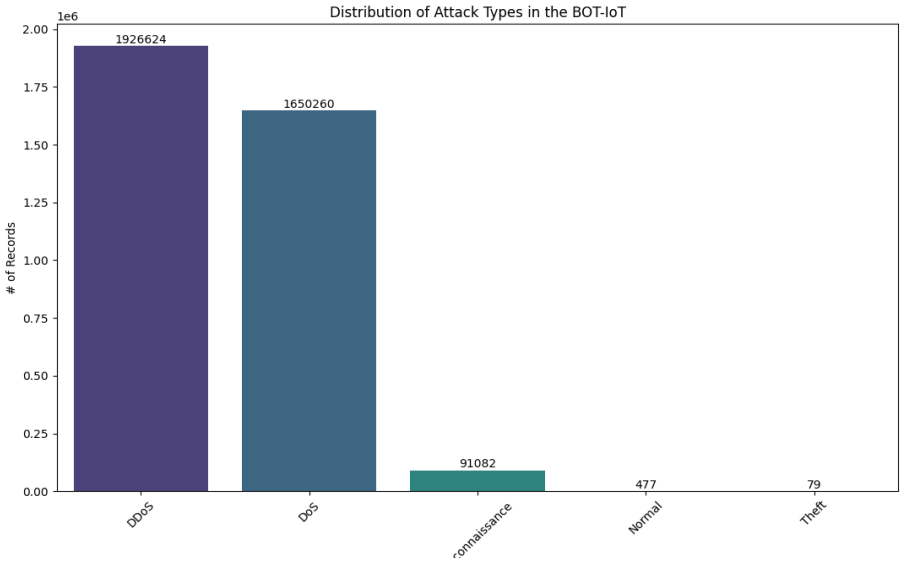
The Bot-IoT dataset, developed by UNSW Canberra's Cyber Range Laboratory in 2020, is recognized as a resource in the field of IoT security. This dataset is designed to simulate a realistic network environment that integrates normal and botnet traffic. It is primarily utilized for the detection and analysis of botnet attacks targeting IoT devices. The dataset is available in various formats, including original pcap files, argus files, and CSV files (Koroniotis et al., 2017, 2019; Koroniotis, Moustafa, Schiliro, et al., 2020; Koroniotis, Moustafa, & Sitnikova, 2020).

The Bot-IoT dataset comprises approximately 1.2 million records, with each record containing 15 distinct features. These records represent various attack scenarios and examples of normal traffic. The dataset was generated in a laboratory setting through a simulated scenario involving the use of multiple IoT devices in a real-world network environment, where botnet attacks were executed to collect data. This approach provides valuable insights into analyzing real-world cyber-attacks.

In the training dataset used in this study, there are 3,668,522 records and 46 features. The dataset is categorized into two primary target labels: normal and attack. The attack labels are further divided into subcategories that specifically represent different types of botnet attacks. The distribution of traffic by main categories is illustrated in Figure 8.

Figure 8

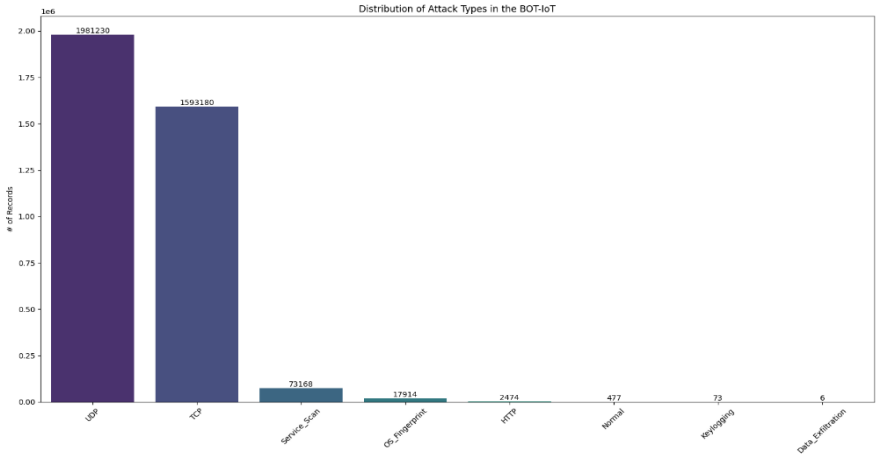
The Distribution of Normal and Attack Traffic in the BoT-IoT Dataset



The Bot-IoT dataset exhibits a significant imbalance, particularly in the representation of normal traffic compared to DDoS and DoS traffic. The volume of normal traffic records is markedly lower, which can adversely affect the model's ability to learn and generalize from the data. Similarly, the category of Theft traffic is also underrepresented, presenting challenges for effective classification. Furthermore, the dataset includes subcategories for the various types of traffic attacks. Figure 9 illustrates the distribution of attack traffic across these subcategories, highlighting the disparities in representation among different attack types. Such imbalances necessitate careful consideration during the training of models, as they may lead to biased learning outcomes favoring the more prevalent categories.

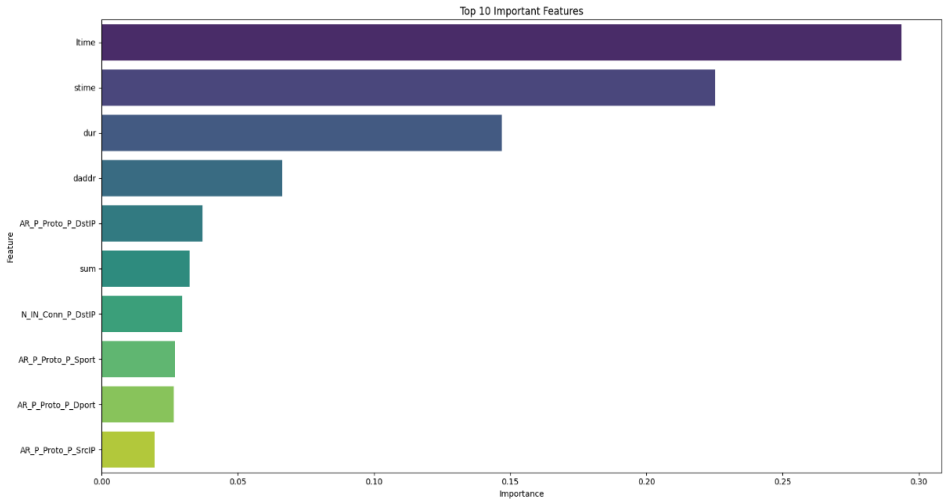
Figure 9

The Distribution of Subcategories in the BoT-IoT Dataset.



Similarly, it can be observed that the subcategories within the dataset do not exhibit a balanced distribution. The scarcity of non-attack normal traffic records is particularly concerning, as this deficiency is likely to adversely impact classification accuracy and increase the rate of false positives. Figure 10 presents the most effective features identified in the classification of the traffic types within the dataset. These features play a crucial role in enhancing the model's performance and its ability to accurately differentiate between normal and attack traffic.

Figure 10
The Top 10 Most Effective Features in Subcategory Classification in BOT-IoT.



In the context of classification, the most effective features identified are the Local Time (ltime) and System Time (stime). The ltime and stime features represent the local time at which an event occurs and the system time, respectively. These features provide critical temporal information regarding the occurrence of events, which is particularly beneficial in time series analyses. The integration of these two timestamps facilitates the synchronization of events between records from different systems, ensuring temporal alignment in the analysis of network traffic or events among IoT devices. Moreover, they play a critical role in time series modeling, enhancing the ability to discern patterns over time.

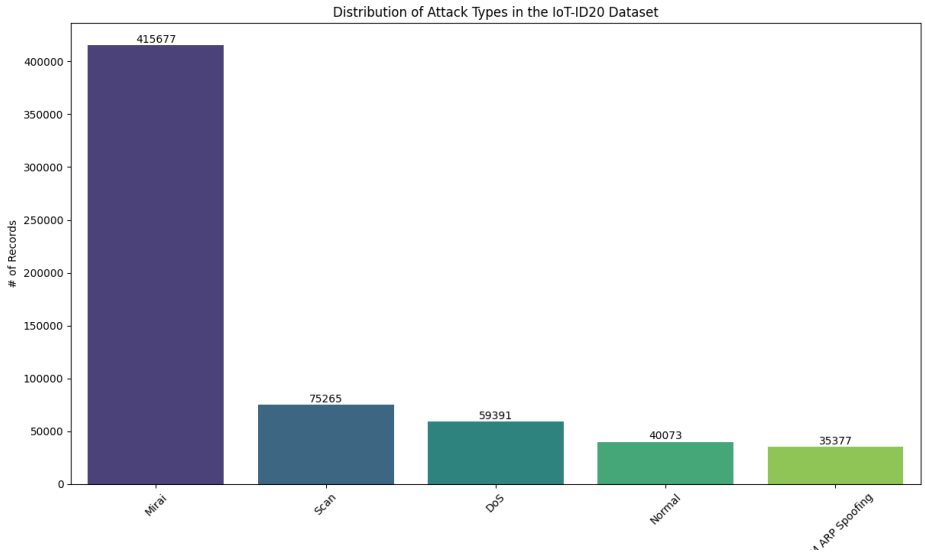
The Bot-IoT dataset is invaluable for research in areas such as machine learning and intrusion detection systems. It is frequently utilized to investigate the diversity and evolution of attacks targeting IoT devices, making it a preferred choice in relevant studies.

IoTID20

This dataset was created in 2020, focusing on attacks against IoT devices and addressing current IoT threats. It has become a frequently preferred resource in recent IoT security research due to its inclusion of up-to-date attack types, making contributions to the field (Surya & Shanthi, 2023).

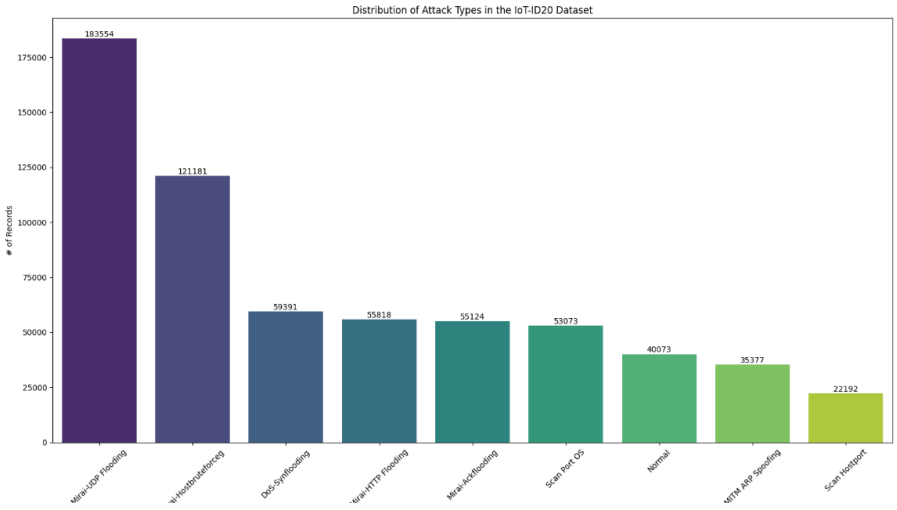
The dataset contains approximately 625,873 records, with a total of 86 different features for each record. The attacks are classified at two levels in the dataset: category and subcategory. The distribution of attack categories and normal traffic is presented in Figure 11.

Figure 11
The Distribution of Traffic Labels in the IoT-ID20 Dataset.



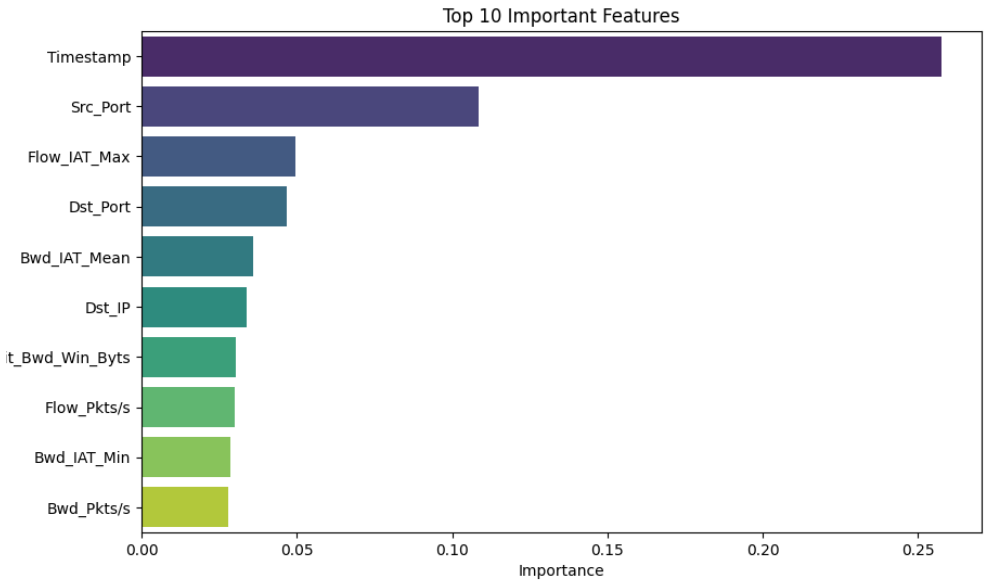
The distribution of the analysis conducted for the subcategories is shown in Figure 12.

Figure 12
The Distribution of Traffic Subcategories in the IoT-ID20 Dataset.



It can be observed that the traffic in both levels of categories is imbalanced. However, it is important that relatively recent attacks, such as the Mirai attack, are included in the dataset. The most effective features are illustrated in Figure 13.

Figure 13
The Top 10 Features in the classification of Attacks

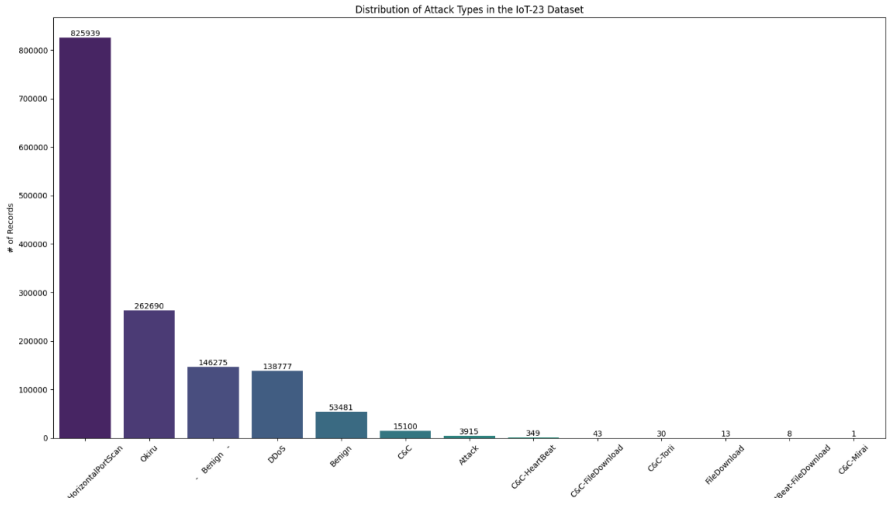


The most influential feature in classification is the timestamp, followed by the source port number.

IoT 23

This dataset has been developed to investigate malicious activities observed in IoT devices (Sharma & Babbar, 2024) (Alfares & Banimelhem, 2024). It contains 1,446,621 records and 28 different features. The dataset was generated in a laboratory environment, where real IoT devices were used to simulate specific attack scenarios and normal communication patterns. The IoT-23 dataset typically exhibits an imbalanced structure. The distribution of traffic within the dataset is illustrated in Figure 14.

Figure 14
The Traffic Classes of the IoT23 Dataset.



The dataset focuses on malware, particularly attacks such as PortScan, Command and Control (C&C), and Distributed Denial of Service attacks. It is utilized for research in

IoT security, with a specific emphasis on malware detection efforts.

General Analysis of Datasets

This section will address the general characteristics and analysis of datasets used in the field of IoT security. Datasets are critical for evaluating the effectiveness of cybersecurity applications and serve as tools for understanding different types of attacks. The differences between balanced and unbalanced datasets are important factors that affect the success of machine learning models. In this context, aspects such as the scope, features, and labeling systems of IoT security datasets will be examined

Balanced datasets are those in which each class has an equal number of examples. In contrast, unbalanced datasets contain some classes with more or fewer examples than others. From the perspective of IoT security, balanced datasets enhance the model's ability to learn each class, while unbalanced datasets may reflect more realistic scenarios.

IoT security datasets vary based on their application areas. Some datasets focus on a specific type of attack, while others include general network traffic or specific devices. For instance, the BoT-IoT dataset targets botnet attacks, whereas the TON_IoT dataset offers a wide range of attack scenarios. Different machine learning techniques may be more effective on different types of datasets. Labeled datasets are used for supervised learning, while less labeled or unlabeled datasets are utilized for unsupervised learning. For example, the NSL-KDD dataset is suitable for supervised learning models.

The sizes of datasets vary based on the number of examples and the number of features they contain. Larger datasets provide opportunities for more comprehensive analyses, while smaller datasets can be processed more quickly. The number of features in IoT security datasets directly impacts model performance. More features enable the model to learn more information, but unnecessary features can increase the model's complexity. Feature engineering plays a crucial role in these datasets. Label types in IoT security datasets can vary. Some datasets contain clear labels indicating a specific type of attack, while others may use more general or ambiguous labels. This situation can affect the accuracy and precision of machine learning models.

Research conducted using IoT datasets provides insights into the field of IoT security. These publications are valuable for identifying new threats, developing new security solutions, and improving existing methods. Table 1 presents a summary view of the datasets addressed in this study.

Table 1

Overview of Various IoT Security Datasets

Dataset	Record Count	Attack Category Count	Features Count	Categories
NSL KDD	125,973	22	42	neptune, satan, ipsweep, portsweep, smurf, nmap, back, teardrop, warezclient, pod, guess_passwd, buffer_overflow, warezmaster, land, imap, rootkit, loadmodule, ftp_write, multihop, phf, perl, spy
CICIDS-2017	2,600,000	14	79	DoS Hulk, PortScan, DDoS, DoS GoldenEye, FTP-Patator, SSH-Patator, DoS slowloris, DoS Slowhttptest, Bot, Web Attack - Brute Force, Web Attack - XSS, Infiltration, Web Attack - Sql Injection, Heartbleed

Dataset	Record Count	Attack Category Count	Features Count	Categories
UNSW-NB15	2,540,044	9	49	Generic, Exploits, Fuzzers, DoS, Reconnaissance, Analysis, Backdoor, Shellcode, Worms
TON-IoT	1,209,013	9	43	Backdoor, DDoS, DoS, Injection, Password, Ransomware, Scanning,XSS, MiTM
BoT-IoT	1,140,000	4	40	DDoS, DoS, Reconnaissance, Theft
IoT-ID20	1,049,045	8	22	Mirai-UDP Flooding, Mirai-Hostbruteforceg, DoS-Synflooding, Mirai-HTTP Flooding, Mirai-Ackflooding, Scan Port OS, MITM ARP Spoofing, Scan Hostport
IoT-23	75,000	12	23	PortScan, Okiru, Benign, DDoS, C&C, Attack, C&C-HeartBeat, C&C-FileDownload, C&C-Torii, FileDownload, C&C-HeartBeat-FileDownload, C&C-Mirai

NSL KDD is a classic intrusion detection dataset that is widely used in the field of cybersecurity. The diversity of categories allows for the learning of various types of attacks during the modeling process. However, the lower number of records compared to other datasets may offer less diversity and a less realistic environment. CICIDS-2017 has a large number of records and represents modern types of attacks. The high number of features allows for the development of more complex and effective models. UNSW-NB15 has a wide number of records and a sufficient set of features, making it suitable for analyzing various attacks. The diversity of attack categories is beneficial for understanding overall security threats. TON-IoT is a dataset specifically designed to examine threats targeting IoT devices. The variety of attacks is crucial for understanding security threats in the IoT environment. BoT-IoT focuses on a specific type of attack and provides data on fundamental IoT threats. However, the limited number of attack categories may restrict the coverage of various attack scenarios. IoT-ID20 represents Mirai-based attacks commonly observed in IoT environments. The specialization of categories is useful for focusing on specific threats. IoT-23, despite having fewer records, includes various attack categories, which is beneficial for understanding different types of attacks.

In general evaluation, CICIDS-2017 and UNSW-NB15 datasets provide a larger number of records, allowing for more comprehensive analyses compared to other datasets. The diversity of attack categories is of great importance for model development and attack detection. Particularly, IoT datasets stand out for containing up-to-date and realistic attack scenarios. Additionally, the high number of features increases the complexity of the dataset and facilitates the creation of more effective machine-learning models.

Conclusion

In this study, we explored the complexities and nuances of IoT security through the analysis of various datasets designed to detect and classify malicious activities. The hypothesis posited that the choice and characteristics of these datasets influence the

performance of machine learning models in identifying security threats within IoT environments. Our findings confirm this hypothesis, demonstrating that datasets with balanced classes and a diverse range of attack types lead to more accurate and reliable models.

The analysis revealed that while datasets like CICIDS-2017 and UNSW-NB15 provide extensive records for training models, the inherent imbalances in certain datasets, such as BoT-IoT and IoT-23, may hinder the detection capabilities of models, particularly for less frequent attack types. Additionally, the effectiveness of certain features, especially time-related attributes, underscores the importance of selecting relevant characteristics that contribute to model performance.

As IoT threats continue to evolve, the integration of synthetic datasets and the continuous refinement of existing datasets will be essential for enhancing the resilience of IoT systems against emerging security challenges. Future research should focus on developing more comprehensive datasets that reflect real-world scenarios, thus enabling researchers to train models that are not only effective but also robust in the face of diverse attack strategies.

In summary, the hypothesis that the characteristics of IoT security datasets impact the efficacy of machine learning models has been substantiated. This study highlights the critical role of dataset selection in advancing IoT security measures and sets the stage for future explorations aimed at fortifying the integrity of IoT environments.

Future efforts could focus on generating datasets with a greater variety of attack types and more balanced classes, which would address the limitations of existing datasets. Synthetic data generation could be further explored to simulate real-world attack scenarios, providing more comprehensive training data for machine learning models.

Further investigation into feature selection techniques could improve model performance by identifying the most relevant features for each attack type. This might involve applying techniques like Principal Component Analysis or Recursive Feature Elimination (RFE) to reduce model complexity while maintaining accuracy. Establishing standardized benchmarks for IoT security datasets would support comparative analysis of models across different datasets. Such benchmarks could help researchers evaluate model robustness and performance more consistently, fostering advancements in the field.

References

Aggarwal, P., & Sharma, S. K. (2015). Analysis of KDD Dataset Attributes—Class wise for Intrusion Detection. *Procedia Computer Science*, *57*, 842–851. https://doi.org/10.1016/j.procs.2015.07.490

Alfares, H., & Banimelhem, O. (2024). Comparative Analysis of Machine Learning Techniques for Handling Imbalance in IoT-23 Dataset for Intrusion Detection Systems. *2024 11th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, 112–119. https://doi.org/10.1109/IOTSMS62296.2024.10710296

Alosaimi, S., & Almutairi, S. M. (2023). An Intrusion Detection System Using BoT-IoT. *Applied Sciences*, *13*(9), 5427. https://doi.org/10.3390/app13095427

Alrayes, F. S., Zakariah, M., Amin, S. U., Iqbal Khan, Z., & Helal, M. (2024). Intrusion Detection in IoT Systems Using Denoising Autoencoder. *IEEE Access*, *12*, 122401–122425. https://doi.org/10.1109/ACCESS.2024.3451726

Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., & Anwar, A. (2020). TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. *IEEE Access*, *8*, 165130–165150. https://doi.org/10.1109/ACCESS.2020.3022862

Ashraf, J., Keshk, M., Moustafa, N., Abdel-Basset, M., Khurshid, H., Bakhshi, A. D., & Mostafa, R. R. (2021). IoTBoT-IDS: A novel statistical learning-enabled botnet detection framework for protecting networks of smart cities. *Sustainable Cities and Society*, *72*, 103041. https://doi.org/10.1016/j.scs.2021.103041

Bansal, K., & Singhrova, A. (2023). Review on intrusion detection system for IoT/IIoT -brief study. *Multimedia Tools and Applications*, *83*(8), 23083–23108. https://doi.org/10.1007/s11042-023-16395-6

Booij, T. M., Chiscop, I., Meeuwissen, E., Moustafa, N., & Hartog, F. T. H. D. (2022). ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets. *IEEE Internet of Things Journal*, *9*(1), 485–496. https://doi.org/10.1109/JIOT.2021.3085194

Guerra, J. L., Catania, C., & Veas, E. (2022). Datasets are not enough: Challenges in labeling network traffic. *Computers & Security*, *120*, 102810. https://doi.org/10.1016/j.cose.2022.102810

Hota, H. S., & Shrivas, A. K. (2014). Decision Tree Techniques Applied on NSL-KDD Data and Its Comparison with Various Feature Selection Techniques. In M. Kumar Kundu, D. P. Mohapatra, A. Konar, & A. Chakraborty (Eds.), *Advanced Computing, Networking and Informatics- Volume 1* (Vol. 27, pp. 205–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-07353-8_24

Karamollaoğlu, H., Doğru, İ. A., & Yücedağ, İ. (2024). An Efficient Deep Learningbased Intrusion Detection System for Internet of Things Networks with Hybrid Feature Reduction and Data Balancing Techniques. *Information Technology and Control*, *53*(1), 243–261. https://doi.org/10.5755/j01.itc.53.1.34933

Kaur, B., Dadkhah, S., Shoeleh, F., Neto, E. C. P., Xiong, P., Iqbal, S., Lamontagne, P., Ray, S., & Ghorbani, A. A. (2023). Internet of Things (IoT) security dataset evolution: Challenges and future directions. *Internet of Things*, *22*, 100780. https://doi.org/10.1016/j.iot.2023.100780

Koroniotis, N., Moustafa, N., Schiliro, F., Gauravaram, P., & Janicke, H. (2020). A Holistic Review of Cybersecurity and Reliability Perspectives in Smart Airports. *IEEE Access*, *8*, 209802–209834. https://doi.org/10.1109/ACCESS.2020.3036728

Koroniotis, N., Moustafa, N., & Sitnikova, E. (2020). A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. *Future Generation Computer Systems*, *110*, 91–106. https://doi.org/10.1016/j.future.2020.03.042

Koroniotis, N., Moustafa, N., Sitnikova, E., & Slay, J. (2017). *Towards Developing Network forensic mechanism for Botnet Activities in the IoT based on Machine Learning Techniques* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1711.02825

Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. *Future Generation Computer Systems*, *100*, 779–796. https://doi.org/10.1016/j.future.2019.05.041

Liang, P., Yang, L., Xiong, Z., Zhang, X., & Liu, G. (2024). Multilevel Intrusion Detection Based on Transformer and Wavelet Transform for IoT Data Security. *IEEE Internet of Things Journal*, *11*(15), 25613–25624. https://doi.org/10.1109/JIOT.2024.3369034

Maiwada, U. D., Imran, S. A., Danyaro, K. U., Janisar, A. A., Salameh, A., & Sarlan, A. B. (2024). Security Concerns of IoT Against DDoS in 5G Systems. *International Journal of Electrical Engineering and Computer Science*, *6*, 98–105. https://doi.

org/10.37394/232027.2024.6.11

Mishra, A. K., Rajput, K., Pandey, N. K., & Pathak, A. (2023). Comparative Analysis of Classification Algorithms Using Bot_IoT Dataset. *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, 1775–1780. https://doi.org/10.1109/ICSCNA58489.2023.10370699

Moustafa, N. (2019). *A Systemic IoT-Fog-Cloud Architecture for Big-Data Analytics and Cyber Security Systems: A Review of Fog Computing* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1906.01055

Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. *Sustainable Cities and Society*, *72*, 102994. https://doi.org/10.1016/j.scs.2021.102994

Moustafa, N., Ahmed, M., & Ahmed, S. (2020). Data Analytics-Enabled Intrusion Detection: Evaluations of ToN_IoT Linux Datasets. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 727–735. https://doi.org/10.1109/TrustCom50675.2020.00100

Moustafa, N., Keshky, M., Debiez, E., & Janicke, H. (2020). Federated TON_IoT Windows Datasets for Evaluating AI-Based Security Applications. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 848–855. https://doi.org/10.1109/TrustCom50675.2020.00114

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). *2015 Military Communications and Information Systems Conference (MilCIS)*, 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

Moustafa, N., & Slay, J. (2016). The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Information Security Journal: A Global Perspective*, *25*(1–3), 18–31. https://doi.org/10.1080/19393555.2015.1125974

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors*, *23*(13), 5941. https://doi.org/10.3390/s23135941

Ozdogan, E. (2024). A Comprehensive Analysis of the Machine Learning Algorithms in IoT IDS Systems. *IEEE Access*, *12*, 46785–46811. https://doi.org/10.1109/ACCESS.2024.3382539

Qing, Y., Liu, X., & Du, Y. (2024). Mitigating data imbalance to improve the generalizability in IoT DDoS detection tasks. *The Journal of Supercomputing*, *80*(7), 9935–9960. https://doi.org/10.1007/s11227-023-05829-5

Saadouni, R., Gherbi, C., Aliouat, Z., Harbi, Y., & Khacha, A. (2024). Intrusion detection systems for IoT based on bio-inspired and machine learning techniques: A systematic review of the literature. *Cluster Computing*, *27*(7), 8655–8681. https://doi.org/10.1007/s10586-024-04388-5

Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization: *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 108–116. https://doi.org/10.5220/0006639801080116

Sharma, A., & Babbar, H. (2024). Understanding IoT-23 Dataset: A Benchmark for IoT Security Analysis. *2023 4th International Conference on Intelligent Technologies (CONIT)*, 1–5. https://doi.org/10.1109/CONIT61985.2024.10627334

Surya, V., & Shanthi, C. (2023). Cross Model Verification of Intrusion Detection

System on IoT Using Convolutional Neural Network. *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 1–12. https://doi.org/10.1109/ICTBIG59752.2023.10456135

Thana-Aksaneekorn, C., Kosolsombat, S., & Luangwiriya, T. (2024). Machine Learning Classification for Intrusion Detection Systems Using the NSL-KDD Dataset. *2024 IEEE International Conference on Cybernetics and Innovations (ICCI)*, 1–6. https://doi.org/10.1109/ICCI60780.2024.10532265

Türk, F. (2023). Analysis of Intrusion Detection Systems in UNSW-NB15 and NSL-KDD Datasets with Machine Learning Algorithms. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *12*(2), 465–477. https://doi.org/10.17798/bitlisfen.1240469

Vibhute, A. D., Patil, C. H., Mane, A. V., & Kale, K. V. (2024). Towards Detection of Network Anomalies using Machine Learning Algorithms on the NSL-KDD Benchmark Datasets. *Procedia Computer Science*, *233*, 960–969. https://doi.org/10.1016/j.procs.2024.03.285

Zafar Iqbal Khan, Mohammad Mazhar Afzal, & Khurram Naim Shamsi. (2024). A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *2*(02), 254–260. https://doi.org/10.47392/IRJAEH.2024.0041

Zoghi, Z., & Serpen, G. (2024). UNSW-NB15 computer security dataset: Analysis through visualization. *SECURITY AND PRIVACY*, *7*(1), e331. https://doi.org/10.1002/spy2.331

About the Authors

Erdal ÖZDOĞAN is an academic affiliated with Management Information Systems, Uludag University, Bursa, Turkiye. He received a Ph.D. degree from the department of Information Systems at Gazi University. He specializes in IoT (Internet of Things), cybersecurity, and network systems. He has a PhD in information sciences and his research interests include IoT, cybersecurity, networks, and artificial intelligence. He has published several papers and book chapters on these topics. He also teaches network security, cryptography, machine learning, and system analysis and design courses.
E-mail: erdalozdogan@uludag.edu.tr , **ORCID:** 0000-0002-3339-0493

Onur CERAN is an academic and researcher affiliated with Gazi University in Türkiye. He received a Ph.D. degree from Gazi University. His work primarily focuses on information security and computer and instructional technologies. He also teaches courses on network, cyber security, ethical hacking, incident response, forensics and IT law courses. He has contributed to various studies and publications on these topics.
E-mail: onur.ceran@gazi.edu.tr, **ORCID:** 0000-0003-2147-0506

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 4%.

The Importance of Dashboard in Data Analysis: An Application Example

M. Hanefi CALP

Ankara Hacı Bayram Veli University

Resul BÜTÜNER

Ministry of National Education

To Cite This Chapter

Calp, M. H., & Bütüner, R. (2024). The Importance of Dashboard in Data Analysis: An Application Example. In M. H. Calp & R. Butuner (Eds.), *Current Studies in Data Science and Analytics* (pp. 144–155). ISRES Publishing.

Introduction

Nowadays, data analysis is of critical importance for businesses, especially in reaching their strategic plans or goals. With the rapid development of the internet and technology, mobile phones and social media applications are widely used. Thus, very large amounts of data/information are produced. (Alan, 2024). Although it is difficult to analyze such large and complex data, it is very important for businesses to use this data in order to make effective and efficient decisions, rather than keeping the data idle in their memories. Today, big data analysis is no longer preferred over traditional methods. At this point, especially in a strong and challenging competitive environment, businesses should use advanced technologies that will effectively perform big data analysis processes in order to gain some advantages over other companies. Companies increase efficiency, reduce costs, and facilitate communication with customers by analyzing the big data they obtain in a short time (Uladi & Arı, 2023; Ayvaz & Salman, 2020). From this perspective, dashboards are among the developed solutions because they facilitate data analysis processes. These panels enable visualization and summary of data obtained through various means. Thus, businesses find an effective analysis opportunity in a much shorter time, increase the speed of access to information, optimize the analysis by monitoring their strategic activities, provide comprehensibility, and preserve data integrity. In addition, dashboards have a simple and informative structure and play an important role in businesses making effective decisions (Yurtay, Ayanoğlu & Yıldız, 2021). In this context, the study aimed to find answers to the extent to which the COVID-19 vaccines produced and used against the virus in the COVID-19 pandemic are related to basic headings such as supply and vaccine inequality. In this context, the data analysis process carried out in the study was supported by dashboards prepared using the Power BI program. Thus, the data will be analyzed more effectively and efficiently in a shorter time. These indicators will also make it easier for institutions, organizations, or individuals to make accurate, efficient, and effective comments and thus make some strategic decisions. In order to achieve the objectives of the study to the maximum extent, the following questions were sought to be answered:

Are states able to deliver vaccines to their citizens equally during the pandemic period?

If not, what are the reasons for this?

What are the effects of vaccine inequality?

How can vaccine inequality be resolved?

What are the parameters used to deliver the vaccine equally to every individual?

The difference and important element of this study from other studies is that it is not only prepared by conducting relevant research, but also provides institutions and organizations with the opportunity and example of implementation with indicator panels, and helps them make strategic decisions by seeing the positive and, if any, negative aspects of the application.

Literature Review

There are many studies in the literature where data is analyzed using data analysis tools such as Power BI, Tableau, etc. and dashboards are created at the end of the analysis process. For example, Picozzi et al. developed 18 key performance indicators for the Clinical Engineering Department of a hospital in Milan and aimed to optimize the maintenance and management of electromedical devices. In this process, an interactive dashboard they created using Power BI helped monitor the maintenance efficiency and obsolescence of the devices. This panel especially supported decision processes. At this point, dashboards were determined in the logistics, technical, and equipment management categories using business intelligence. This dashboard provided a comprehensive framework for continuous monitoring and decision-making processes. The results showed that the developed KPIs and dashboard have a high potential to increase operational insights and improve the maintenance processes of the healthcare facility (Picozzi, Nocco, Pezzillo, De Cosmo & Cimolin, 2024).

Gonçalves et al. evaluated the effects of business intelligence tools on decision-making processes in organizations, especially in the sales and marketing field. The study revealed that business intelligence systems allow faster and more effective analysis by implementing key performance indicators (KPIs) with data integration and transformation. As a result, it was emphasized that data-integrated dashboards play an important role in the decision-making process (Gonçalves, Gonçalves & Campante, 2023).

In their studies, Khilari et al. discussed the role and importance of business intelligence tools such as Power BI in performance management. The study discussed how these tools can improve business processes by providing more effective information to decision-makers thanks to their real-time data analysis and visualization capabilities. In addition, it was emphasized that Power BI facilitates the data analysis process with its user-friendly interface and various dashboards. As a result, it was revealed that business intelligence tools such as Power BI and Tableau play an important role in data analysis and visualization processes. It was concluded that these tools help users effectively manage their data sets and achieve business goals. It was also stated that Power BI is more suitable for small data sets and Tableau is more suitable for large data sets (Khilari, Singh & Mane, 2022).

Larasati et al. In the study, they investigated the implementation of a business intelligence dashboard (dashboard) for the BRIN Technology Services Center Public Service Institution. The study analyzed the financial data of the last five years and revealed the relationships between revenue realization and expenditures. It was emphasized how this information can support public service decision-making processes. In addition, the potential of business intelligence tools to improve performance in the public sector was also discussed. The result of the study showed that the business intelligence dashboard was an effective tool to improve the financial performance of the BRIN Technology Services Center. It was emphasized that this dashboard provided the necessary insights

for leaders to make better decisions and that improvements should be made in receivables management. It was also concluded that more care should be taken in the selection of technology service partners. Thus, it was aimed to increase the cash balance (Larasati, Tanzil, Alfian & Wardani, 2024).

Antal et al. In this study, they demonstrated the use of business intelligence software in the mining sector, improving the management of mining mechanization through data analysis and dashboards. The study aimed to provide an effective tool for monitoring performance indicators and supporting decision-making processes by analyzing failure data through the Power BI application. Using the Microsoft Power BI application, the researchers carried out the stages of collecting, processing, and visualizing data. Thus, interactive dashboards were created to monitor performance indicators and support decision-making processes. The study concluded that business intelligence software makes significant contributions to the analysis of processes and monitoring performance indicators in the mining sector. The created dashboards allow managers to focus on critical areas and make improvements thanks to the effective visualization and analysis of failure data (Antal, Marasova, Hájıček, Klapko & Mitrik, 2022).

Qi and Nagalingham emphasized the importance of Business Intelligence tools in analyzing and visualizing health data by addressing the rise of diabetes in the United States. In particular, the use of tools such as Tableau aimed to develop strategies for predicting and preventing diabetes by better-analyzing diseases and lifestyles. In addition, the effects of education and income levels on health were examined. The conclusion of the study was that business intelligence solutions allow for more effective management of diabetes through the analysis and visualization of health data. BI tools have improved the processes of monitoring and preventing diseases by providing healthcare managers with the opportunity to make more informed decisions. As a result, it has been shown that such solutions help organizations improve healthcare services and achieve better patient outcomes (Qi & Nagalingham, 2023).

Aprillia et al. In the study, they discussed the design of a dashboard to monitor the distribution of government aid in Jepara. The research aims to facilitate the visualization and analysis of data by switching from traditional reporting methods (table format) to a more user-friendly interface. The designed dashboards include various features that allow users to monitor aid distribution more effectively. The result of the study showed that the designed dashboard significantly improved the system of monitoring government aid distribution in Jepara. The dashboard presents data through graphical visualization, allowing users to understand and evaluate performance indicators more easily. As a result, this system has increased the effectiveness of aid distribution by supporting decision-making processes (Aprillia, Noranita, Kom & Tech, 2021).

Awamleh et al. In the study, they examined the impact of international performance indicators on sustainable development and the role of business intelligence techniques in this process. The research emphasizes ways to achieve sustainability goals using organizational agility and data science applications and reveals the benefits of sustainable applications in social, economic, and environmental dimensions. In addition, the importance of strategic partnerships for sustainable development is emphasized. The result of the study showed that organizational agility when combined with business intelligence systems and data science applications, has a positive impact on sustainable development. This integration supports economic, social, and environmental development by strengthening strategic decision-making processes and facilitates organizations to achieve their sustainability goals (Awamleh, Alarabiat, & Bustami, 2024).

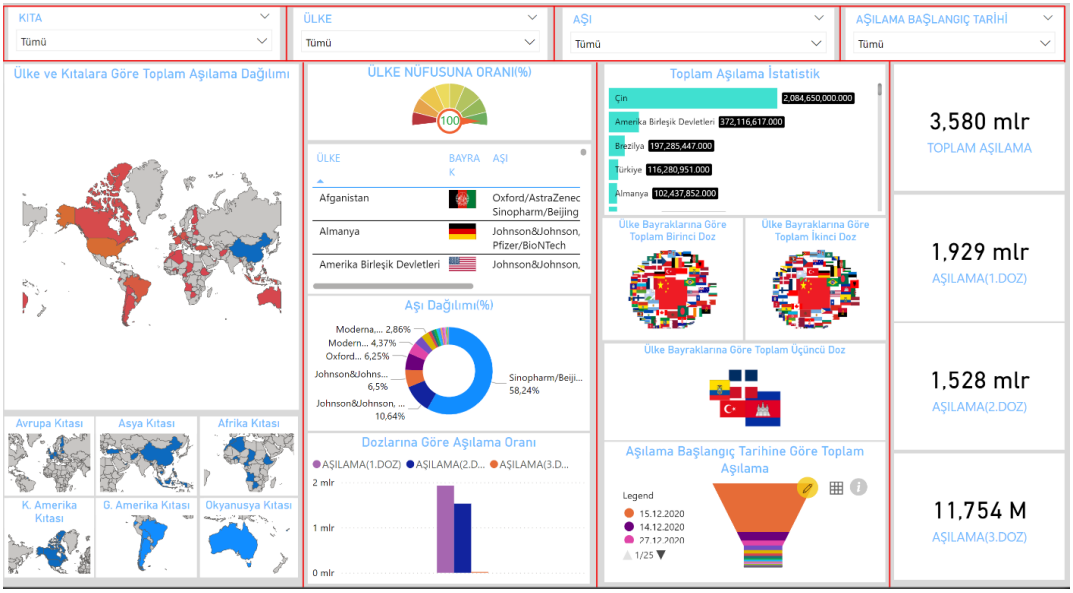
Arnaboldi et al. examined how self-service business intelligence tools, especially dashboards, used in university administration transform performance management processes. The study showed that these tools accelerate decision-making processes by facilitating users' access to data and reducing managers' dependency on data integration. In addition, dashboards emphasized the changes in organizational dynamics and the

positive feedback of users towards these tools. The results of the study showed that self-service business intelligence tools, especially dashboards, create a significant transformation in accounting and organizational processes. These tools increase users' interaction with data, causing accountants to question their traditional roles and open the door to organizational innovations. As a result, the adoption of new technologies brings with it complex dynamics such as organizational shocks and users' increasing interest in data (Arnaboldi, Robbiani, & Carlucci, 2021).

An Application Example

This section includes all the details of an application example prepared to demonstrate the importance of dashboards, which are an important step in data analysis. Power BI software was used for data analysis and dashboards. Each module is visualized one by one with dashboards and its functions are explained. The general dashboard of the application is given in Figure 1.

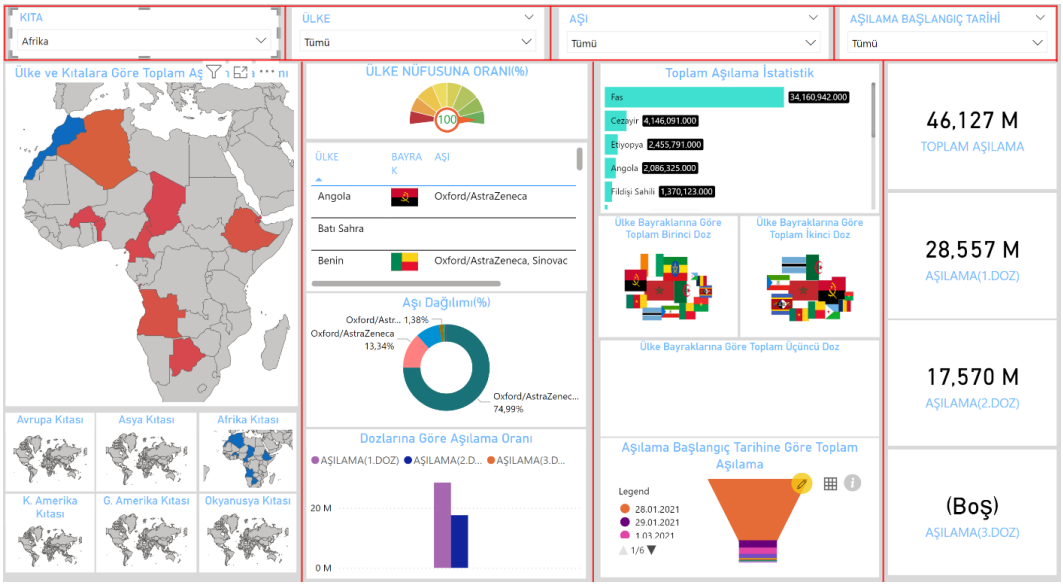
Figure 1
General view of the dashboard



Continent Module

With the continent module, information such as the total number of vaccinations in which continent, the names of the vaccines administered, the doses administered, which countries in the continent have had their first, second, and third dose vaccination studies, the location of the continent on the world map, the countries in the relevant continent, the date on which the most vaccinations were made in the continent, which country in the continent has made the most total vaccinations, etc. can be viewed by the user on the relevant dashboard.

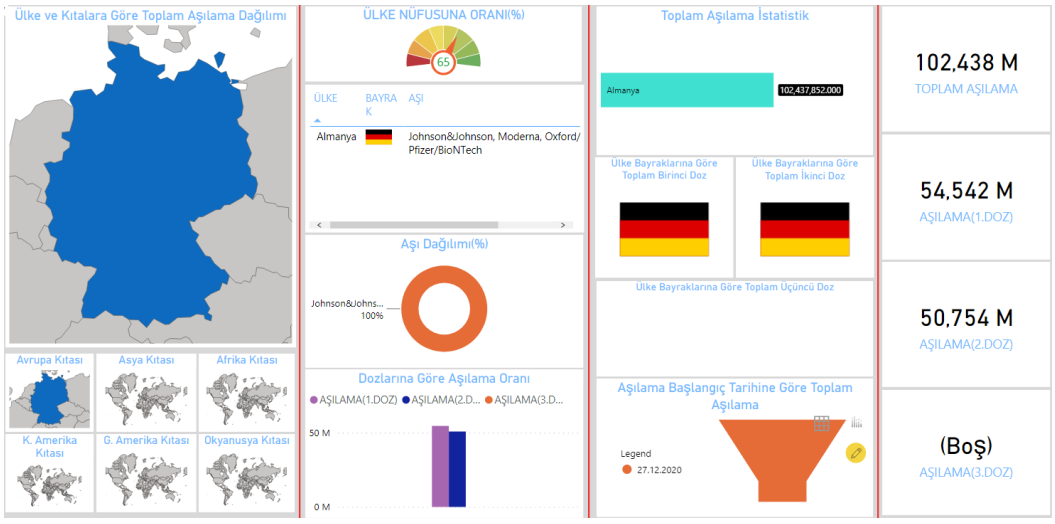
Figure 2
Relevant data visuals for the African continent



Country Module

With the country module, information such as the total number of vaccinations in each country, the names of the vaccines administered in the country, the doses administered, the first, second, and third dose vaccination studies conducted by the country, the ratio of the total vaccinations conducted by the country to the country's population, the continent in which the country is located, the country's location on the world map, and the country's vaccination start date can be viewed on the user's relevant dashboard.

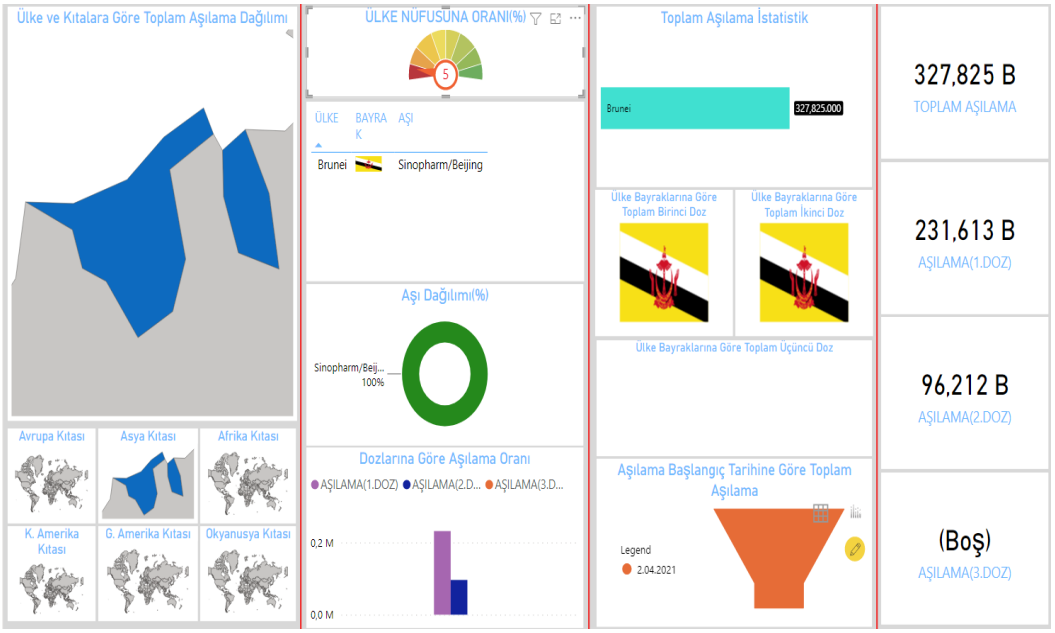
Figure 3
Relevant data visuals for the selected country



Vaccination Module

With the vaccination module, the total number of vaccinations from which vaccines, the names of the vaccines administered, the number of doses administered from which vaccines, the ratio of the vaccine administered to the country's population, which vaccine was used in which continent or continents, which vaccine was used in which country or countries, and in which country the relevant vaccine group was used the most on which date, can be seen on the dashboard by the user by selecting the relevant country.

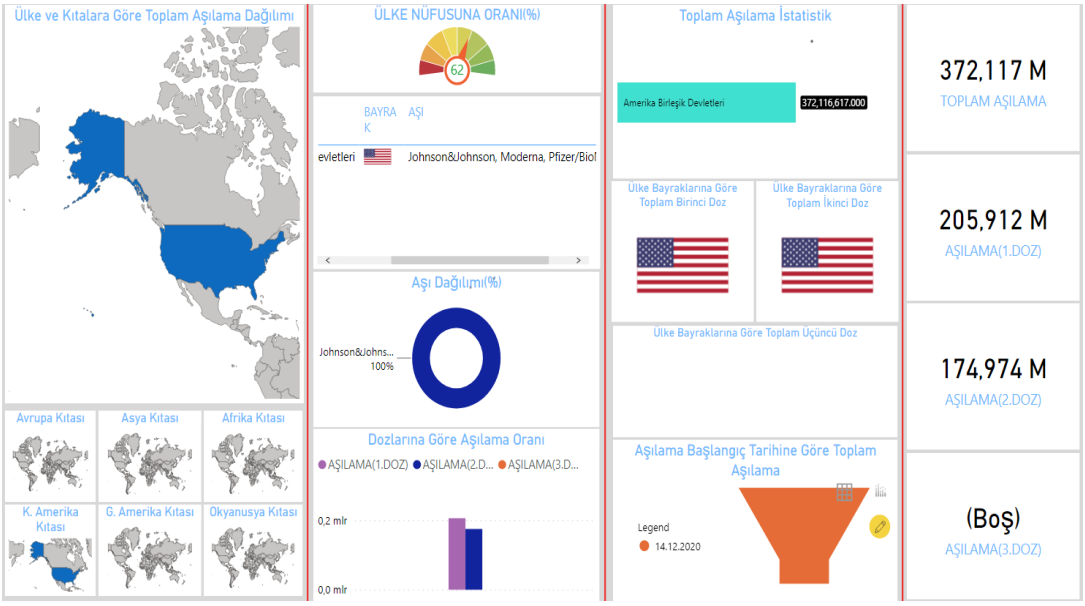
Figure 4
Relevant data visuals for the selected vaccine



Vaccination Start Date Module

With the vaccination start date module, one of the dates that vaccination started is selected and the country or countries that started vaccination on the relevant date, which continent or continents were studied on these dates, how much vaccination was done in total on the relevant date, the dose study applied on the relevant date, the ratio of the vaccine administered on the relevant date to the country's population, which vaccines were used on the selected date, can be seen on the dashboard by the user selecting the relevant date.

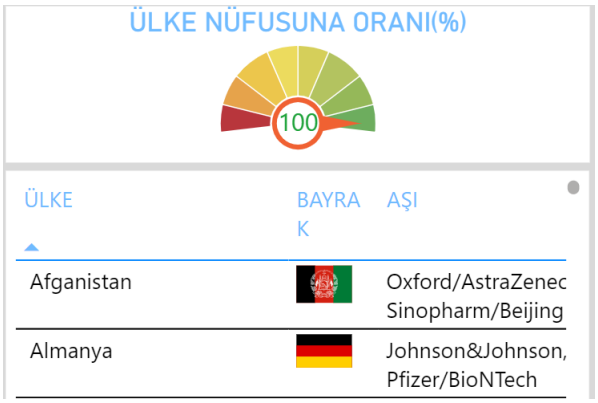
Figure 5
Relevant data visuals of the selected vaccine



Country Population Ratio Module

With this module, the vaccination status of the country selected from the country module is calculated starting from the vaccination start date and transferred to the user, and the results can be seen on the dashboard.

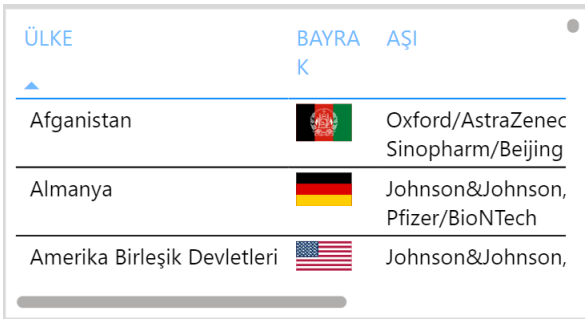
Figure 6
Indicator of how much of the country's population has been vaccinated since the vaccination start date of the selected country



Country, Flag, and Vaccine Module

With this module, the countries belonging to the continent selected from the continent module, the country selected from the country module, the country or countries using the vaccine selected from the vaccine module, the country or countries starting on the relevant date selected from the vaccination start date module, the flag image of the countries and detailed vaccines used are transferred to the user for all of these.

Figure 7
Countries belonging to the selected continent, flag, and vaccine information

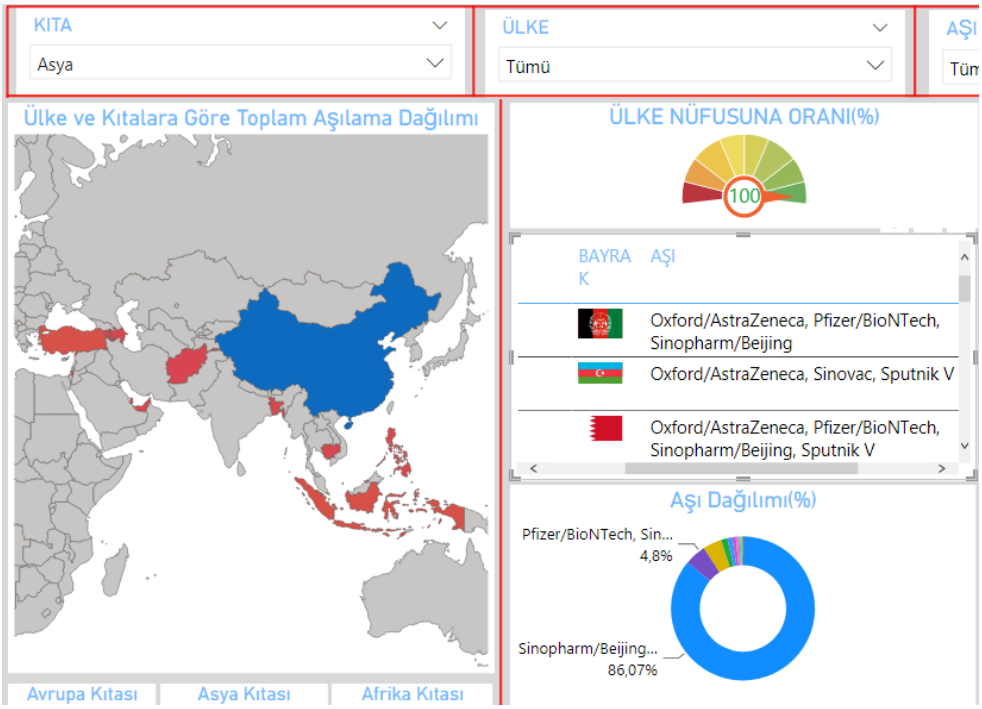


Vaccine Distribution Module

This module shows the user the rates of vaccines or vaccines used in the country, the rates of vaccines applied on the selected continent, and the rates at which vaccines are used in the world.

Figure 8

Vaccination rates for the relevant continent are shown to the user

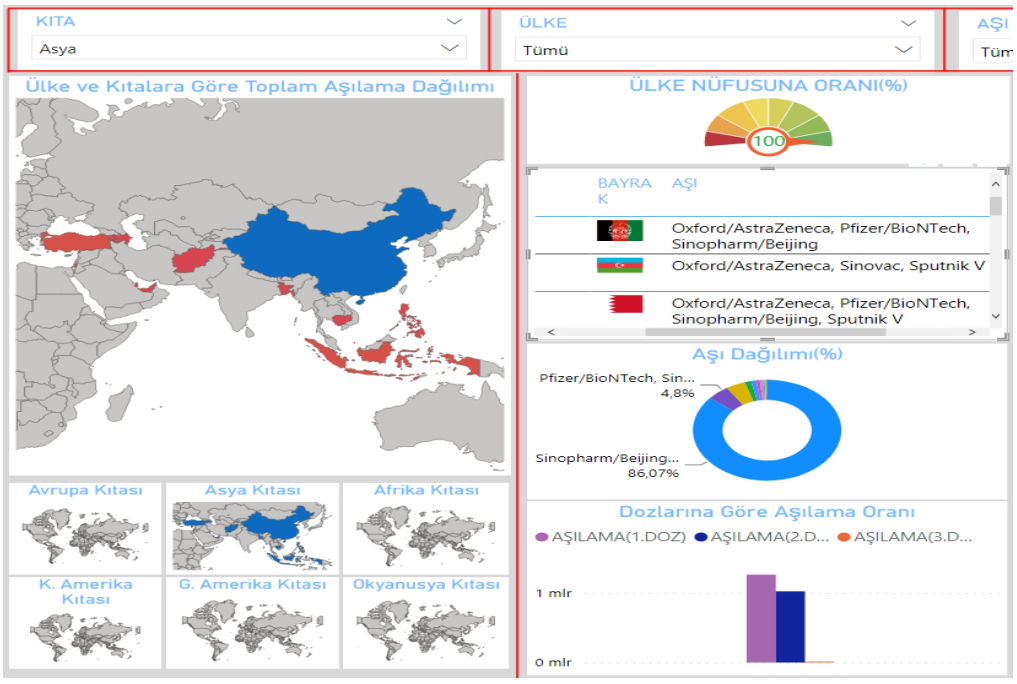


Vaccination Rate Module According to Doses

With this module, dose usage information is displayed to the user with the help of graphics for the first, second, and third doses applied within the country, on the continent, and around the world.

Figure 9

Vaccination rate according to doses for the selected country

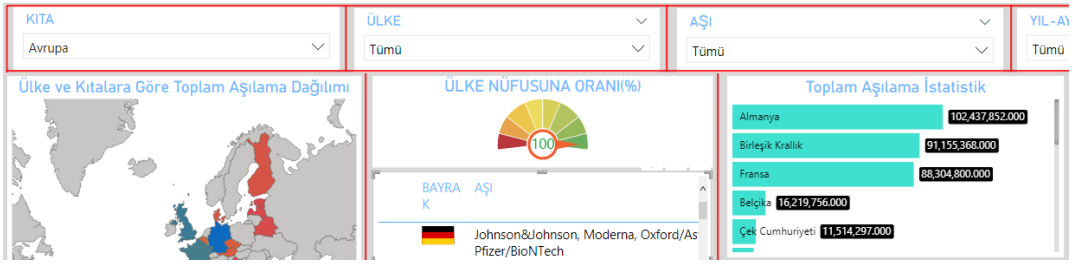


Total Vaccination Statistics Module

This module transfers the total vaccination (first, second, third dose) on the selected continent and the entire world from most to least to the user. It transfers the countries where the selected vaccine is high and in which it is low from the vaccination module to the user. By selecting the date from the vaccination start date, the order of the countries that have applied the most and least vaccinations on the relevant date is displayed.

Figure 10

Total vaccination data for the selected continent



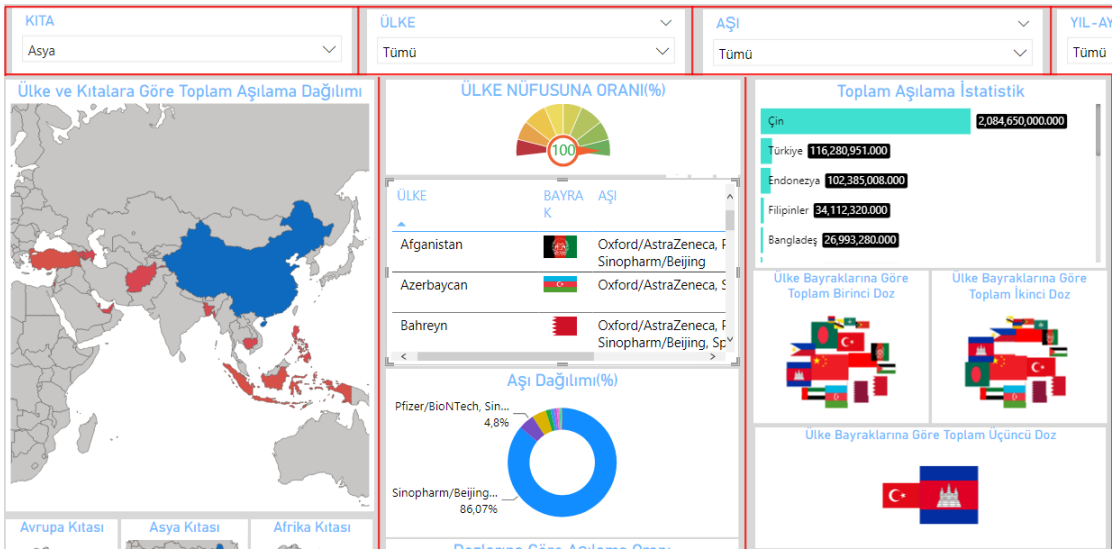
Total vaccination data of countries all over the world is transferred to the user in order from most to least. A vaccine group is selected from the vaccine module and information about which countries use more and which countries use less vaccines is transferred to the user. The date is selected from the vaccination start date and the countries with the most and least vaccinations for this date are seen.

Dose Module by Country Flag

With this module, the selection is made according to the parameters of continent, country, vaccine, and vaccination start date. The flags of the countries that applied the first, second, and third doses are included in the relevant module according to the dose applied. The flag of the country is included in the module according to the doses applied for the country. According to the relevant vaccination start date parameter, the country or countries whose vaccination start date is the selected date come, and their flags are included in the relevant module, and the applied dose information is transferred to the user.

Figure 11

Doses of the countries belonging to the selected continent according to their flags

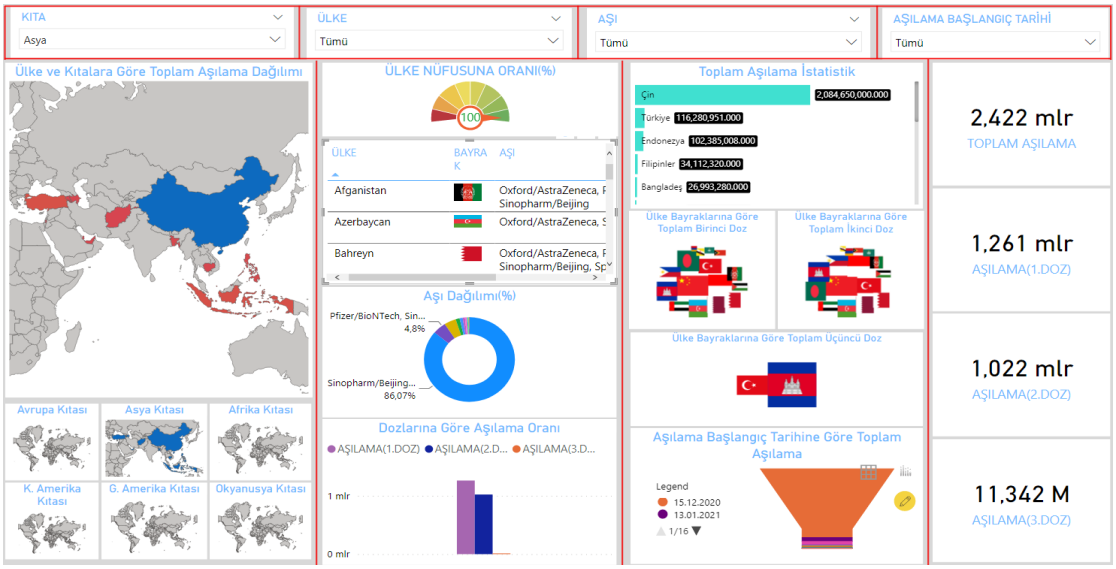


Total Vaccination Module According to Vaccination Start Date

This module provides the user with the total vaccination information of the countries, in order of the date on which the most vaccinations were made, from the most to the least. The total vaccinations for the relevant continent are listed in order of the date on which the most vaccinations were applied. For example, in Figure 12, it is seen that the total vaccination in the Asian continent was done in China on 15.12.2020. It is learned that the total vaccinations for the relevant vaccine or vaccines were done in which country and on which date. It is seen that the total vaccinations for the relevant vaccine were applied in Germany on 27.12.2020.

Figure 11

Total vaccination module according to vaccination start date

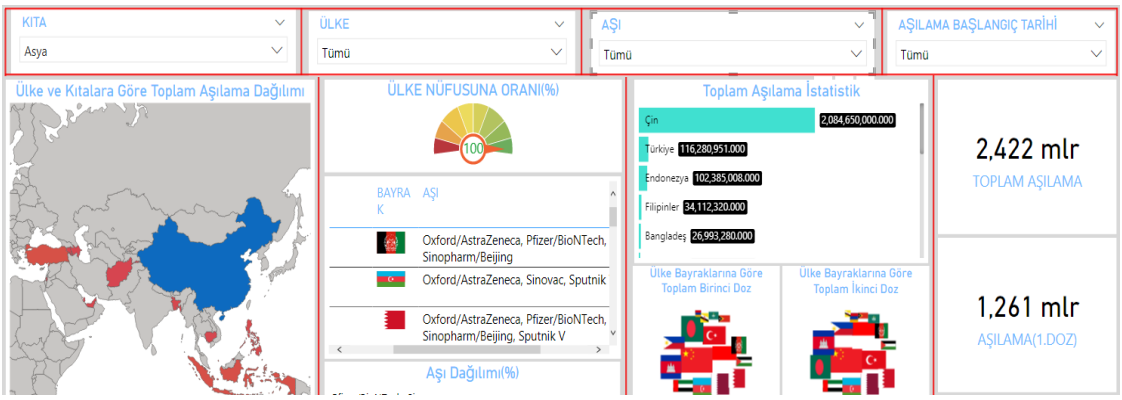


Total Vaccination Module

This module uses our country, continent vaccine, and vaccination start date selection modules, and the total number of people vaccinated in relation to this selected parameter data is transferred to the user. The relevant vaccination start date is selected. Total vaccination information is transferred to the user depending on the selected vaccination start date.

Figure 12

Total vaccination information for the selected country



Conclusion and Recommendations

In this study, the importance of dashboards in data analysis processes has been demonstrated through an application example. The fact that safe and effective COVID-19 vaccines have a great social value has increased the motivation of the study. At this point, the development and production of COVID-19 vaccines specified in the Health Policy document, their affordability, allocation, and distribution, as well as the effects of vaccine inequality and the absence of vaccine inequality were analyzed using the Power BI analysis application.

The dashboards used revealed that vaccine inequality, although complex, is still a solvable problem. However, the defining characteristics of vaccine diplomacy and its potential effects on COVID-19 immunization were examined in the light of vaccine empathy. The results underlined the instrumental and indispensable role of vaccine diplomacy in solving the problem of vaccine inequality in the midst of the pandemic. Finally, with the application example, it was proven that dashboards can be used effectively and efficiently in the analysis process of a health problem, thus clearly demonstrating the importance of dashboards.

References

Alan, S. (2024). Sosyal medya paylaşım ve yorumlarının konaklama işletmelerinde stratejik yönetim kararlarına etkisi üzerine bir araştırma. Necmettin Erbakan Üniversitesi, Sosyal Bilimler Enstitüsü Turizm İşletmeciliği Anabilim Dalı, Konya.

Antal, R., Marasova, D., Hájıček, R., Klapko, P., & Mitrik, V. (2022). Implementation of business intelligence system to analyze the data for mining mechanization-case study. Acta Montanistica Slovaca, 27(3).

Aprillia, D. A., Noranita, B., Kom, M., & Tech, M. I. Dashboard Design as Monitoring System Distribution of Government Assistance.

Arnaboldi, M., Robbiani, A., & Carlucci, P. (2021). On the relevance of self-service business intelligence to university management. Journal of Accounting & Organizational Change, 17(1), 5-22

Awamleh, F. T., Alarabiat, Y. A., & Bustami, A. N. (2024). Enhancing sustainable development through international performance indicators: The role of business intelligence techniques. Chall. Sustain, 12(3), 203-218.

Ayvaz, S., Salman, Y. B. (2020). Türkiye'de Firmaların Büyük Veri Teknolojileri Bilinirliği ve Kullanımı Analizi. Avrupa Bilim ve Teknoloji Dergisi, (18), 728-737.

Gonçalves, C. T., Gonçalves, M. J. A., & Campante, M. I. (2023). Developing Integrated Performance Dashboards Visualisations Using Power BI as a Platform. Information, 14(11), 614.

Khilari, D. S., Singh, C., & Mane, M. B. (2022). Business Intelligence Tool-Power BI for Performance Management. Available at SSRN 4177482.

Larasati, D., Tanzil, N. D., Alfian, A., & Wardani, L. (2024). Business Intelligence Dashboard for Financial Performance Analysis of Public Service Agency Using Microsoft Power BI. JASa (Jurnal Akuntansi, Audit dan Sistem Informasi Akuntansi), 8(2), 491-499.

Picozzi, P., Nocco, U., Pezzillo, A., De Cosmo, A., & Cimolin, V. (2024). The Use of Business Intelligence Software to Monitor Key Performance Indicators (KPIs) for the Evaluation of a Computerized Maintenance Management System (CMMS).

Electronics, 13(12), 2286.

Qi, S. S. J., & Nagalingham, S. (2023). Business Intelligence Data Visualization for Diabetes Health Prediction. International Journal of Advanced Computer Science and Applications, 14(1).

Uladi, A. İ., & Arı, E. S. (2023). Büyük Veri, Büyük Veri Analizi ve Uygulama Alanları. Yönetim Bilişim Sistemleri Dergisi, 9(1), 1-14.

Yurtay, Y., Ayanoğlu, M., & Yıldız, T. K. (2021). Gösterge Panelinin Üretim Bağlamındaki Karar Süreçlerine Etkisi Üzerine Bir Ampirik Çalışma. Kocaeli Üniversitesi Sosyal Bilimler Dergisi, 1(41), 1-15.

About the Authors

Assoc. Prof. Dr. M. Hanefi CALP received Ph.D. degree from the department of Management Information Systems at Gazi University, one of the most prestigious universities in Türkiye. He works as an Associate Professor in the Department of Management Information Systems of the Faculty of Economics & Administrative Sciences of the Ankara Hacı Bayram Veli University. His research interest includes Management Information Systems, Digital Transformation, Artificial Neural Networks, Expert Systems, Fuzzy Logic, Risk Management, Risk Analysis, Human-Computer Interaction, Technology Management, Knowledge Management, and Project Management.
E-mail: hanefi.calp@hbv.edu.tr , **ORCID:** 0000-0001-7991-438X

Engineer, MSc. Resul BÜTÜNER is an information technology teacher at the Ministry of National Education in Ankara, Türkiye. He has a master's degree in Computer Engineering from Necmettin Erbakan University. He is currently working on a book in the field of artificial intelligence, robotic coding, data mining, and augmented reality applications. He is an instructor in the field of Robotic coding within TUBITAK. He continues to write a book in the field of robotic coding at the Ministry of National Education. He worked as a coordinator in projects related to student education. .
E-mail: resul.butuner@eba.gov.tr , **ORCID:**0000-0002-9778-2349

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 3%.

The book "*Current Studies in Data Science and Analytics*" consists of selected topics invited by the editors.

This edition includes chapters on Data Science, Data Analytics, Big Data and the Internet of Things etc. used in today's technology.
The aim of the book is to provide readers with an opportunity for academic peer-reviewed publication in the fields of data science, data analytics of data-driven methodologies.

Current Studies in Data Science and Analytics is published by ISRES Publications.

ISRES Publishing