Hesaplamalı Zekanın Kuramsal Temelleri: Yapay Zeka, Öğrenme Kuramı ve Büyük Veri Paradigması

Editör: Doç. Dr. Atınç Yılmaz

ÖZGÜR YAYINLARI

Hesaplamalı Zekanın Kuramsal Temelleri: Yapay Zeka, Öğrenme Kuramı ve Büyük Veri Paradigması

Editör:

Doç. Dr. Atınç Yılmaz

ÖZGÜR YAYINLARI

Published by
Özgür Yayın-Dağıtım Co. Ltd.
Certificate Number: 45503

15 Temmuz Mah. 148136. Sk. No: 9 Şehitkamil/Gaziantep
+90.850 260 09 97
+90.532 289 82 15
www.ozguryayınlari.com
info@ozguryayinlari.com

Hesaplamalı Zekanın Kuramsal Temelleri: Yapay Zeka, Öğrenme Kuramı ve Büyük Veri Paradigması

Editör: Doç. Dr. Atınç Yılmaz

Language: Turkish-English
Publication Date: 2026
Cover design by Mehmet Çakır

Print and digital versions typeset by Çizgi Medya Co. Ltd.

ISBN (PDF): 978-625-8813-27-2

DOI: https://doi.org/10.58830/ozgur.pub1351

Suggested citation:
Yılmaz, A. (ed) (2026). *Hesaplamalı Zekanın Kuramsal Temelleri: Yapay Zeka, Öğrenme Kuramı ve Büyük Veri Paradigması.* Özgür Publications. DOI: https://doi.org/10.58830/ozgur.pub1351. License: CC-BY-NC 4.0

The full text of this book has been peer-reviewed to ensure high academic standards. For full review policies, see https://www.ozguryayinlari.com/

ÖZGÜR YAYINLARI

Ön Söz

Yapay zeka, son yetmiş yıl içerisinde yalnızca bilgisayar bilimlerinin bir alt araştırma alanı olmaktan çıkmış; ekonomi, sağlık, eğitim, mühendislik, güvenlik ve sosyal bilimler başta olmak üzere insan yaşamının hemen her alanını dönüştüren disiplinlerarası bir paradigma haline gelmiştir. Özellikle veri üretim hızındaki artış, hesaplama gücünün gelişmesi ve öğrenme algoritmalarının olgunlaşması, yapay zekâyı teorik bir araştırma alanından günlük yaşamın merkezinde yer alan bir teknolojiye dönüştürmüştür.

Ancak yapay zeka alanında yaşanan hızlı gelişmeler, çoğu zaman bu teknolojilerin altında yatan kuramsal temellerin göz ardı edilmesine neden olmaktadır. Oysa günümüzde kullanılan makine öğrenmesi, derin öğrenme, büyük veri analitiği ve karar destek sistemleri; matematiksel modelleme, hesaplama kuramı, istatistiksel öğrenme ve optimizasyon gibi köklü bilimsel birikimlerin üzerine inşa edilmiştir. Bu nedenle yapay zekayı yalnızca uygulama boyutuyla değerlendirmek, alanın bütüncül olarak anlaşılmasını engelleyen önemli bir eksiklik oluşturmaktadır.

Çalışma, yapay zeka, öğrenme kuramı ve büyük veri paradigmasının kuramsal temellerini, güncel uygulama alanlarıyla birlikte ele alan disiplinlerarası bir başvuru kaynağı olarak hazırlanmıştır. Kitapta, öğrenen sistemlerin matematiksel ve hesaplamalı altyapısı incelenirken; sağlık, siber güvenlik, sosyal medya analitiği, çok-kipli veri işleme ve veri kıtlığı altında öğrenme gibi güncel araştırma alanlarında geliştirilen yöntemler de ayrıntılı biçimde değerlendirilmektedir. Böylece yapay zekâ, yalnızca algoritmik bir yaklaşım olarak değil, büyük ölçekli veriden bilgi üreten ve farklı disiplinlerde dönüşüm yaratan bütüncül bir bilimsel paradigma olarak ele alınmaktadır. Nitekim büyük veri paradigması, günümüz yapay zeka sistemlerinin gelişimini hızlandıran ve öğrenme süreçlerini yeniden tanımlayan temel itici güçlerden biri haline gelmiştir.

Bu eser, yalnızca yapay zeka teknolojilerini kullanan bireylere değil; bu teknolojilerin neden ve nasıl çalıştığını anlamak isteyen araştırmacılara, lisans ve lisansüstü öğrencilerine, akademisyenlere ve sektör profesyonellerine de rehberlik etmeyi amaçlamaktadır. Kitabın temel hedefi, okuyucunun yapay zekâya ilişkin kavramsal ve kuramsal altyapısını güçlendirmek, güncel gelişmeleri daha sağlıklı yorumlayabilmesini sağlamak ve gelecekte ortaya

çıkacak teknolojik dönüşümlere bilimsel bir perspektifle yaklaşmasına katkıda bulunmaktır.

Yapay zekanın giderek daha fazla otonom karar aldığı, öğrenebildiği ve insan yaşamını şekillendirdiği bir dönemde, bu alanın kuramsal temellerini anlamanın yalnızca akademik bir gereklilik değil, aynı zamanda toplumsal bir sorumluluk olduğu düşünülmektedir. Bu kitabın, yapay zeka, öğrenme kuramı ve büyük veri alanlarında çalışan araştırmacılar, öğrenciler ve sektör profesyonelleri için güvenilir bir başvuru kaynağı olması ve gelecekteki disiplinlerarası araştırmalara ilham vermesi amaçlanmaktadır.

İçindekiler

Chapter 1

Artificial Intelligence-Based Cyberbullying Detection and Prevention: Deep Learning Architectures, Multimodal Analysis, Ethical Challenges, and Future Directions

Sara Naghib Zadeh[1]

Zühre Aydin[2]

Abstract

The rapid growth of social media and digital communication platforms has significantly increased the prevalence of cyberbullying, online harassment, and hate speech. Due to the large volume and dynamic nature of online content, manual monitoring has become insufficient, leading to the growing use of artificial intelligence (AI)-based detection and prevention systems. Cyberbullying is not only a technical problem but also a major social and psychological challenge with serious consequences for individuals and online communities.

This paper presents a comprehensive review of AI-based cyberbullying detection approaches, focusing on machine learning, deep learning, and multimodal analysis techniques. The study examines traditional machine learning methods alongside advanced deep learning architectures such as CNN, RNN, LSTM, hybrid CNN–LSTM models, and transformer-based models including BERT. In addition, the paper discusses multimodal systems that combine textual, visual, and sentiment-based analysis to improve the detection of implicit and complex harmful content.

The study also addresses important challenges such as adversarial attacks, linguistic manipulation, dataset imbalance, algorithmic bias, privacy concerns, and ethical issues related to automated moderation systems. Furthermore,

1 Dr. Lecture, Halic University, Vocational School, Department of Computer Programming, ORCID: 0009-0005- 6959-1165.

2 Halic University, Vocational School, Department of Big Data Analytics, ORCID: 0009-0009-4523-9406

https://doi.org/10.58830/ozgur.pub1351.c5534

future directions involving explainable AI, predictive moderation systems, and human–AI collaborative frameworks are explored.

The findings indicate that although AI-based systems have significantly improved cyberbullying detection performance, achieving a balance between technical accuracy, fairness, transparency, and freedom of expression remains a major challenge. Future progress in this field will require interdisciplinary approaches that integrate advanced AI technologies with ethical and humancentered moderation strategies.

1. Cyberbullying Development in the Digital Environment and Detection Approaches

With the rapid expansion of communication technologies and the pervasive influence of social media in everyday life, human interaction has increasingly shifted from physical environments to digital ecosystems. This transformation has fundamentally changed the nature of social communication, enabling instant information exchange, global connectivity, and unprecedented access to digital platforms. However, alongside these advantages, the same environment has also facilitated the emergence and rapid spread of harmful online behaviors such as cyberbullying, online harassment, and hate speech.

One of the most critical factors contributing to cyberbullying is anonymity in online environments. Anonymity reduces accountability and psychological inhibition, allowing individuals to express aggressive behaviors that they would typically avoid in face-to-face interactions. Research has consistently shown that anonymity plays a central role in increasing the likelihood and severity of online aggression . In addition, the perceived distance between users in digital environments further amplifies disinhibition effects, making cyberbullying more frequent and less controllable compared to traditional bullying (Al-Ajlan & Ykhlef, 2018; Al-Dabet et al., 2023).

From a psychological and social perspective, cyberbullying has significant and long-lasting consequences, especially among adolescents. This group is particularly vulnerable due to their developmental stage, higher dependency on peer validation, and intensive use of social media platforms. Exposure to repeated online harassment can lead to serious mental health issues such as anxiety, depression, reduced self-esteem, social withdrawal, and in extreme cases, suicidal ideation .Unlike traditional bullying, the digital nature of cyberbullying ensures continuous exposure, as harmful content can persist online indefinitely and be accessed repeatedly (Aldreabi, 2024; Alabdulwahab et al., 2023).

Furthermore, cyberbullying is not an isolated individual behavior but a complex socio-technical phenomenon influenced by platform design, user

interaction patterns, and algorithmic content distribution. Recommendation systems, content virality mechanisms, and social reinforcement loops can unintentionally amplify harmful content, thereby increasing its visibility and impact. As illustrated in Figure 1, the rapid expansion of digital communication environments, combined with anonymity and large-scale social interaction, has accelerated the spread of cyberbullying and increased the need for AI-based automated detection systems.

In response to these challenges, automated detection systems based on machine learning have been introduced as a fundamental solution for identifying harmful content in large-scale social media environments. These systems are capable of processing massive volumes of textual data and detecting linguistic patterns associated with offensive, abusive, or threatening content. Studies demonstrate that machine learning algorithms perform effectively in classifying cyberbullyingrelated content, particularly in platforms such as Twitter where data volume and velocity are high (Muneer & Fati, 2020).

The integration of Natural Language Processing (NLP) techniques with machine learning further enhances system performance by enabling more accurate feature extraction from textual data. NLP techniques such as tokenization, sentiment analysis, and word embedding representations play a crucial role in improving classification accuracy and contextual understanding (LeCun et al., 2015).

Despite their effectiveness, traditional machine learning approaches such as Support Vector Machines (SVM), Naïve Bayes, and Random Forest have inherent limitations. These models rely heavily on manual feature engineering and often fail to capture contextual dependencies, sarcasm, and semantic complexity in language. As a result, their performance decreases significantly in realworld social media environments where language is dynamic, informal, and context-dependent (Aqeel & Kamble, 2022).

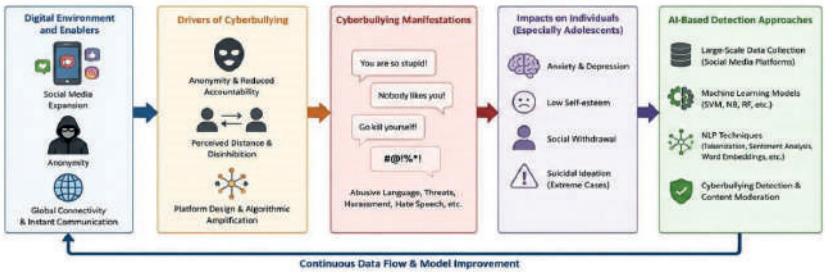


Figure 1. Evolution of Cyberbullying in Digital Environments and AI-Based Detection Framework

2. Deep Learning Architectures for Cyberbullying Detection

Deep learning represents a significant advancement in machine learning, enabling the modeling of complex nonlinear relationships through multi-layer neural network architectures. These models are capable of automatically extracting hierarchical representations from raw data, eliminating the need for manual feature engineering and significantly improving classification performance in complex tasks (LeCun et al., 2015).

In the field of cyberbullying detection, deep learning has become the dominant methodological approach due to its ability to handle large-scale, unstructured, and noisy textual data. Table 1 provides a comparative overview of the most commonly used deep learning architectures in cyberbullying detection, highlighting their strengths, limitations, and application domains.

2.1. CNN-Based Models

Convolutional Neural Networks (CNNs) are widely used for text classification tasks due to their ability to capture local patterns and spatial relationships between words. In cyberbullying detection, CNNs are particularly effective in analyzing short and informal texts commonly found on social media platforms. These models can identify local semantic patterns such as offensive phrases, repeated word structures, and contextual word groupings. Wang et al. demonstrate that CNN-based models achieve high accuracy in detecting abusive and offensive language in short social media texts (Wang et al., 2014).

2.2. RNN and LSTM-Based Models

Recurrent Neural Networks (RNNs) are designed to capture sequential dependencies in text data, making them suitable for modeling sentence-level context. However, traditional RNNs suffer from vanishing gradient problems, which limit their ability to learn long-term dependencies. To address this limitation, Long Short-Term Memory (LSTM) networks were introduced, enabling more effective learning of long-range contextual relationships in text sequences (Meta Transparency Center, 2024; Aldreabi & Blackburn, 2023).

LSTM-based models are particularly effective in detecting subtle linguistic cues such as implicit aggression and contextual negativity, which are often overlooked by simpler models.

2.3. Hybrid Architectures

To leverage the strengths of both CNN and RNN architectures, hybrid models such as CNN–LSTM and CNN–LRCN have been proposed. These

models combine local feature extraction (CNN) with sequential learning (RNN/LSTM), resulting in improved classification performance. Empirical studies show that hybrid architectures significantly outperform traditional machine learning methods in cyberbullying and toxic content detection tasks (Meta Transparency Center, 2024; SJ & Cho, 2020).

Hybrid models are particularly effective in handling social media data, which is typically short, noisy, and linguistically inconsistent.

2.4. Transformer-Based Models

The introduction of transformer architectures has revolutionized natural language processing. Models such as BERT (Bidirectional Encoder Representations from Transformers) provide deep contextual understanding by analyzing text in both forward and backward directions simultaneously. This bidirectional learning mechanism enables more accurate semantic representation of language (European Union, 2018).

Pre-trained BERT models have shown strong performance in detecting cyberbullying, hate speech, and offensive content. These models are especially effective in handling complex linguistic phenomena such as sarcasm, irony, and implicit aggression, which are challenging for traditional architectures (DataTurks, 2018; Mozafari et al., 2019).

Table 1. Comparison of Deep Learning Architectures for Cyberbullying Detection

Architecture Type	Key Idea	Strengths	Limitations	Application in Cyberbullying Detection	Representative References
CNN-Based Models	Extract local features using convolution filters over text	Efficient feature extraction, strong for short texts, captures local patterns (e.g., offensive phrases)	Limited ability to model long-range dependencies	Detects abusive words, offensive phrases, and short toxic posts in social media	[109]
RNN / LSTM-Based Models	Models sequential dependencies in text data	Captures context and temporal dependencies, effective for sentence-level understanding	Vanishing gradient (RNN), higher computational cost (LSTM)	Detects implicit aggression, contextual negativity, and sequential linguistic cues	[31], [52]
Hybrid Models (CNN–LSTM, CNN–LRCN)	Combines CNN for feature extraction and RNN/LSTM for sequence modeling	Higher accuracy, captures both local and global context	More complex architecture, higher training cost	Effective for noisy social media text and mixed linguistic patterns	[31], [97]

Transformer-Based Models (e.g., BERT)	Uses self-attention mechanism for bidirectional context learning	Captures deep contextual meaning, handles sarcasm and implicit hate speech well	Computationally expensive, requires large-scale pretraining	State-of-the-art performance in cyberbullying, hate speech, and offensive language detection	[17], [73], [58]

3. AI-Based Security Systems and Future Perspectives in Cyberbullying Prevention

The rapid expansion of digital communication platforms has created new challenges for ensuring user safety in online environments. Social media platforms, messaging applications, and online communities generate vast amounts of user-generated content every day, making manual monitoring increasingly difficult. As a result, artificial intelligence (AI)-based security systems have emerged as essential tools for supporting content moderation, risk assessment, and cyberbullying prevention efforts (Hussain et al., 2018; Wang et al., 2021).

AI-based security systems are designed to identify, analyze, and respond to harmful online behaviors in real time. Unlike traditional moderation approaches that rely heavily on human reviewers, these systems can continuously monitor large-scale digital environments and assist platforms in detecting potentially harmful interactions more efficiently. Automated moderation tools help reduce response times and improve the scalability of online safety operations (Aqeel & Kamble, 2022).

One of the most important developments in this area is the integration of real-time content moderation mechanisms. Modern security frameworks can automatically flag suspicious content, prioritize high-risk interactions for human review, and support platform administrators in enforcing community guidelines. Such systems contribute to creating safer online environments while reducing the workload of moderation teams (Hussain et al., 2018).

Beyond content moderation, AI technologies are increasingly being used for behavioral analysis. Rather than focusing solely on individual messages, advanced security systems can evaluate patterns of user behavior, interaction frequency, communication networks, and historical activity records. This broader perspective enables the identification of recurring harmful behaviors and potential risk factors associated with cyberbullying incidents (Wang et al., 2021).

Future security frameworks are expected to become more predictive rather than purely reactive. Predictive intelligence systems aim to identify behavioral indicators that may signal the emergence of harmful online interactions before significant damage occurs. By combining behavioral analytics, social network analysis, and large-scale data processing, these systems may provide early warning mechanisms for cyberbullying prevention (Jahan & Oussalah, 2023; Aqeel & Kamble, 2022).

Another important trend is the development of multimodal AI systems. Online communication is no longer limited to text-based interactions. Images, videos, emojis, GIFs, voice recordings, and multimedia content increasingly shape user communication patterns. Consequently, future cyberbullying prevention frameworks are expected to incorporate multiple data sources to achieve a more comprehensive understanding of online behavior. Studies indicate that multimodal analysis can improve the identification of harmful content that may not be explicitly expressed through text alone (Wang et al., 2022).

Large-scale social media platforms also require adaptive security infrastructures capable of responding to rapidly changing communication patterns. New forms of online harassment continuously emerge through evolving slang, coded language, and platform-specific behaviors. Therefore, future AI security systems must be capable of continuous learning and adaptation to maintain effectiveness in dynamic digital environments (Weimann & Masri, 2020).

Despite these advancements, the deployment of AI-based security systems raises significant ethical and regulatory concerns. Automated moderation systems may produce biased outcomes, disproportionately affect certain user groups, or incorrectly restrict legitimate forms of expression. Questions regarding transparency, accountability, privacy protection, and freedom of speech remain central challenges in the development of trustworthy AI systems (Floridi & Cowls, 2019; Hosseini et al., 2017).

To address these concerns, researchers increasingly advocate the adoption of explainable AI principles. Explainable systems provide greater transparency regarding how moderation decisions are made, allowing users, platform administrators, and policymakers to better understand automated outcomes. Such transparency is essential for building trust and ensuring responsible AI governance (Hosseini et al., 2017; Selbst et al., 2019).

Several Explainable AI (XAI) techniques have been proposed to improve transparency in cyberbullying detection systems. Methods such as SHAP

(SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) help identify the most influential features behind model predictions. In transformer-based architectures, attention-based explanation methods further enhance interpretability by highlighting words and contextual patterns that contribute to classification outcomes. These techniques improve transparency, fairness assessment, error analysis, and user trust in AI-driven moderation systems (Arrieta et al., 2020; Aldreabi & Blackburn, 2023).

Looking ahead, AI-based security systems are expected to evolve into integrated digital safety ecosystems that combine real-time monitoring, predictive risk assessment, multimodal analysis, and human oversight. These systems will play a critical role in supporting cyberbullying prevention strategies while balancing technological effectiveness with ethical responsibility and user rights (Diaz-Garcia & Carvalho, 2024; Wang et al., 2021).

Large Language Models (LLMs) and foundation models have recently emerged as powerful tools for cyberbullying detection and content moderation. Models such as GPT, Llama, PaLM, and Claude are capable of understanding complex linguistic structures, contextual meaning, implicit aggression, sarcasm, and culturally dependent expressions. Unlike traditional machine learning systems, LLMs can perform zero-shot and few-shot classification, reducing the need for taskspecific training datasets. Furthermore, foundation models can support multilingual moderation and adaptive content analysis across different online platforms. However, concerns regarding hallucinations, bias, explainability, computational cost, and ethical governance remain significant challenges for their large-scale deployment in cyberbullying prevention systems (Brown et al., 2020; Diaz-Garcia & Carvalho, 2024; Ziems et al., 2023).

Table 2. Emerging AI Approaches in Cyberbullying Prevention

Approach	Main Benefit	Key Challenge
Predictive AI	Early risk detection	False predictions
Multimodal AI	Detects text and visual abuse	High complexity
Explainable AI (XAI)	Improves transparency and trust	Limited interpretability
Large Language Models (LLMs)	Advanced contextual understanding	Bias and ethical concerns

4. Impacts of Cyberbullying on Individuals and Society, and Data and Ethical Challenges

Cyberbullying is not merely a technical issue but a complex social and psychological phenomenon with far-reaching consequences. At the individual level, victims of cyberbullying often experience psychological stress, anxiety, depression, reduced self-esteem, and social isolation. These effects are particularly severe among adolescents due to their developmental vulnerability and high exposure to social media platforms (Aldreabi, 2024; Al-Ajlan & Ykhlef, 2018; Al-Dabet et al., 2023; Alabdulwahab et al., 2023).

One of the most critical characteristics of cyberbullying is the persistence of digital content. Unlike offline bullying, harmful content in digital environments can remain accessible indefinitely, leading to repeated exposure and prolonged psychological harm (Vidgen & Yasseri, 2020; Hussain et al., 2018). In addition, perpetrators of cyberbullying may also experience negative consequences, including reduced empathy and normalization of aggressive behavior over time (Aljalaoud et al., 2022).

At the societal level, cyberbullying contributes to reduced trust in digital platforms, increased polarization, and decreased social participation. The bystander effect further exacerbates the issue, as individuals who witness online aggression often fail to intervene (Al-Dabet et al., 2023; Erliyani, 2021; Gomez et al., 2020).

From a technical perspective, the effectiveness of detection systems depends heavily on dataset quality. Data scarcity, imbalance, and multilingual variability remain significant challenges (DiazGarcia & Carvalho, 2024; Selbst et al., 2019). In addition, real-world social media data is often noisy and unstructured, which complicates model training (Ashraf et al., 2023; Wulczyn et al., 2017).

Ethical challenges also play a critical role in system design. Key concerns include privacy protection, false positive detection, algorithmic bias, and balancing moderation with freedom of expression (Hosseini et al., 2017; Jahan & Oussalah, 2023; Weidinger et al., 2022). These issues highlight the need for responsible AI development in this field.

Multimodal systems have shown improved performance compared to text-only models, especially in detecting meme-based and visually embedded harmful content (MohammedJany et al., 2023; Weimann & Masri, 2020; Zampieri et al., 2019). However, their computational complexity and interpretability remain ongoing challenges. Table 3 summarizes the individual, societal, technical, and ethical impacts of cyberbullying, along with key challenges in AI-based detection systems.

Table 3. Impacts of Cyberbullying and Key Technical & Ethical Challenges in Detection Systems

Category	Sub-Domain	Description	Key Issues
Individual Impact	Psychological effects	Anxiety, depression, low selfesteem, social isolation	High vulnerability in adolescents
Individual Impact	Behavioral effects	Reduced empathy, normalization of aggression	Long-term behavioral changes
Content Characteristics	Persistence of content	Harmful content remains online indefinitely	Repeated exposure and psychological harm
Societal Impact	Social consequences	Reduced trust, polarization, low participation	Weak bystander intervention
Data Challenges	Data quality	Scarcity, imbalance, multilingual variability	Poor generalization of models
Data Challenges	Data structure	Noisy, unstructured social media data	Training difficulty
Ethical Challenges	Privacy & fairness	Data protection, algorithmic bias	Risk of unfair decisions
Ethical Challenges	Moderation balance	Freedom of expression vs control	Over/under moderation risk
Technical Advancement	Multimodal AI	Text + image + video detection	High complexity, low interpretability

5. AI-Based Preventive Strategies for Cyberbullying Mitigation

Modern approaches to cyberbullying mitigation increasingly focus on prevention rather than reaction. AI-based systems are now designed to detect early warning signals and prevent harmful interactions before they escalate (Aldreabi & Blackburn, 2023; Hussain et al., 2018).

Machine learning models analyze behavioral features such as writing tone, interaction frequency, and temporal changes in user activity to identify potential risk patterns (Biggio et al., 2012; LeCun et al., 2015). Deep learning models with temporal and attention mechanisms further enhance prediction accuracy by capturing behavioral evolution over time (Hochreiter & Schmidhuber, 1997).

Real-time content moderation systems are widely used in social media platforms to filter harmful content instantly (DataTurks, 2018; Aldreabi &

Blackburn, 2023). However, these systems struggle with indirect expressions such as sarcasm and implicit aggression (McMahan et al., 2017; Weimann & Masri, 2020).

Transformer-based models such as BERT significantly improve contextual understanding in preventive systems (Devlin et al., 2019; Mozafari et al., 2019). Additionally, multilingual models enhance the ability to detect cyberbullying across different languages and cultural contexts (Muneer & Fati, 2020; Ullah et al., 2022).

Hybrid human-AI moderation systems are also widely adopted. In these systems, AI performs initial detection while human moderators handle ambiguous or sensitive cases (Troop-Gordon et al., 2019; Weidinger et al., 2022). This improves both accuracy and fairness.

However, challenges such as algorithmic bias, adversarial manipulation, and over-censorship remain significant issues (Hosseini et al., 2017; Selbst et al., 2019; Papernot et al., 2016). Therefore, future research focuses on developing explainable and ethical AI systems.

6. Integration of Humans, Education, and Systems in Cyberbullying Mitigation

Effective cyberbullying mitigation requires a combination of technological, educational, and legal strategies. Digital literacy plays a crucial role in reducing user vulnerability and promoting safe online behavior (Al-Hashedi et al., 2022; Alabdulwahab et al., 2023).

Human-AI collaboration is essential because fully automated systems often struggle with contextual understanding. Hybrid systems combining machine learning and human supervision provide higher reliability and accuracy (Troop-Gordon et al., 2019; Aldreabi & Blackburn, 2023).

Human moderators are particularly important for interpreting ambiguous content such as sarcasm, humor, and culturally specific expressions (Mozafari et al., 2019; Wang et al., 2022). This reduces misclassification errors and improves fairness (Weidinger et al., 2022).

At the policy level, regulatory frameworks are increasingly being implemented to ensure platform accountability and user protection (Erliyani, 2021; Aljalaoud et al., 2022). However, these regulations must balance security with privacy and freedom of expression (Weidinger et al., 2022; Selbst et al., 2019).

7. Cyberbullying Detection Using Machine Learning and Deep Learning Techniques

The rapid growth of social media platforms and online communication environments has significantly increased both the volume and complexity of cyberbullying-related content. Platforms such as Twitter, Instagram, Facebook, Reddit, TikTok, and YouTube generate massive amounts of user-generated content every day, making manual moderation impractical and inefficient (Aldreabi, 2024; Al-Dabet et al., 2023). Consequently, automated cyberbullying detection systems based on machine learning and deep learning have become essential tools for maintaining safer online environments.

Traditional machine learning approaches were among the earliest methods applied to cyberbullying detection. These systems typically rely on handcrafted linguistic features, including bag-of-words representations, n-grams, term frequency–inverse document frequency (TF-IDF), sentiment scores, and syntactic patterns. Algorithms such as Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, and Random Forest have shown reasonable performance in offensive language classification tasks (Al-Dabet et al., 2023; Aqeel & Kamble, 2022).

Despite their effectiveness in structured environments, traditional machine learning methods face significant limitations in real-world social media contexts. Online communication often includes slang expressions, abbreviations, intentional misspellings, sarcasm, emojis, and evolving linguistic patterns that are difficult to capture through manually engineered features alone. As a result, these systems frequently struggle to interpret contextual meaning and semantic dependencies between words (LeCun et al., 2015).

Deep learning approaches have addressed many of these limitations by enabling automatic feature extraction directly from raw textual data. Neural network architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks can learn hierarchical representations of language without requiring extensive manual preprocessing (Meta Transparency Center, 2024; Aldreabi & Blackburn, 2023).

CNN-based architectures are particularly effective in identifying local semantic patterns within short social media texts. These models can detect repeated offensive structures, contextual word combinations, and abusive linguistic patterns with high accuracy (Wang et al., 2014). In contrast, RNN and LSTM models focus on sequential dependencies and contextual flow,

making them more suitable for understanding sentence-level semantics and long-range contextual relationships.

Hybrid architectures combining CNN and LSTM have demonstrated superior performance in cyberbullying detection tasks by integrating local feature extraction with sequential contextual modeling (Meta Transparency Center, 2024). Such models are particularly useful in noisy and unstructured social media environments where linguistic information is fragmented and contextdependent (SJ & Cho, 2020).

In recent years, transformer-based architectures such as BERT have significantly improved the performance of cyberbullying detection systems. Unlike earlier models, transformers use attention mechanisms to capture contextual relationships between words more effectively. BERT-based models are capable of understanding implicit aggression, hate speech, sarcasm, and contextsensitive language patterns that are difficult for traditional models to interpret (European Union, 2018; Mozafari et al., 2019).

Recent advances in Foundation Models and Large Language Models (LLMs) have introduced new possibilities for cyberbullying detection and content moderation. Models such as GPT-based systems, LLaMA, PaLM, and other large-scale transformer architectures are capable of understanding contextual meaning, implicit aggression, sarcasm, and nuanced harmful language with greater accuracy than traditional classifiers (Brown et al., 2020; Diaz-Garcia & Carvalho, 2024). Unlike task-specific models, LLMs can perform zero-shot and few-shot classification, reducing the need for extensive labeled datasets and enabling more flexible moderation across diverse online environments (Brown et al., 2020; Ziems et al., 2023). Furthermore, LLM-based moderation systems can provide contextual reasoning and assist human moderators in complex decision-making processes, improving the detection of subtle and context-dependent forms of harmful content (Diaz-Garcia & Carvalho, 2024; Ziems et al., 2023). However, concerns regarding hallucinations, bias propagation, computational cost, privacy, transparency, and accountability remain important challenges for their deployment in large-scale moderation environments (Weidinger et al., 2022; OpenAI, 2024).

Another major challenge in this field is multilingual cyberbullying detection. Social media platforms contain content in multiple languages, dialects, and writing styles. Studies indicate that models trained on one language often perform poorly when applied to other linguistic environments, highlighting the need for multilingual and cross-cultural approaches (Alabdulwahab et al., 2023; Muneer & Fati, 2020).

Datasets also play a critical role in the effectiveness of cyberbullying detection systems. Public datasets such as "Cyber Trolls," Twitter hate speech datasets, and Reddit comment collections are widely used for model training and benchmarking (Aqeel & Kamble, 2022; Ashraf et al., 2023). However, dataset imbalance remains a major challenge because harmful content often represents only a small fraction of total online communication. This imbalance can lead to biased classification systems and increased false negative rates (Selbst et al., 2019).

Moreover, annotation inconsistency presents another significant issue. The interpretation of cyberbullying often depends on cultural context, social norms, and subjective judgment, making reliable labeling difficult. Consequently, researchers increasingly emphasize the importance of balanced, diverse, and context-aware datasets (Xu et al., 2012).

Despite substantial progress, cyberbullying detection systems are still imperfect. False positives may incorrectly classify harmless content as abusive, while false negatives may fail to identify harmful interactions. Such errors can negatively affect both user trust and platform credibility (Weidinger et al., 2022). Therefore, current research increasingly focuses on explainable and fair AI systems that provide more transparent and accountable decision-making processes (Hosseini et al., 2017; Selbst et al., 2019). Table 4 provides a comprehensive comparison of traditional machine learning and deep learning approaches for cyberbullying detection, highlighting their key features, strengths, limitations, and application domains.

Table 4. Comparison of Machine Learning and Deep Learning Approaches for Cyberbullying Detection

Category	Methods / Models	Key Features	Strengths	Limitations	Application in Cyberbullying Detection
Traditional Machine Learning	SVM, Naïve Bayes, Logistic Regression, Random Forest	Handcrafted features (TFIDF, n-grams, sentiment, syntax)	Fast, interpretable, low computational cost	Poor contextual understanding, limited scalability	Basic offensive language and spam detection
CNN-Based Deep Learning	Convolutional Neural Networks	Local feature extraction from text	Strong performance on short texts, detects local patterns	Weak in longterm dependencies	Detection of abusive phrases in social media
RNN / LSTM Models	Recurrent Neural Networks, LSTM	Sequential dependency modeling	Captures context and semantic flow	Vanishing gradient, higher training cost	Sentence-level cyberbullying detection

Hybrid Models	CNN–LSTM, CNN–LRCN	Combines spatial + sequential learning	Higher accuracy, robust to noisy data	Complex architecture, computational cost	Advanced social media text classification
Transformer-Based Models	BERT and variants	Attention-based contextual learning	Strong semantic understanding, handles sarcasm	Requires large data and resources	State-of-the-art cyberbullying detection
Multilingual Challenges	Cross-lingual models	Languageagnostic representation	Enables multilingual detection	Performance drop across languages	Global social media analysis
Dataset Issues	Imbalanced, noisy datasets	Real-world social media data	Reflects real conditions	Bias, imbalance, annotation inconsistency	Training limitation for all models

8. Multimodal Approaches and Sentiment-Based Cyberbullying Analysis

Cyberbullying has extended beyond text-based communication to include images, videos, memes, emojis, and audio content. Therefore, text-only detection systems are no longer sufficient for effective identification of harmful behavior (Ashraf et al., 2023).

Multimodal systems integrate visual and textual information using computer vision models (e.g., CNNs) and NLP-based architectures such as RNN, BiLSTM, and transformers (Atoum, 2021). This integration allows better contextual understanding, especially when offensive meaning arises from the combination of text and images, such as memes with implicit aggression (Weimann & Masri, 2020).

Traditional NLP models often fail to capture such implicit forms of cyberbullying, whereas multimodal approaches significantly improve classification accuracy by combining multiple information sources (Wang et al., 2022).

Sentiment analysis further enhances detection by identifying emotional tone, sarcasm, irony, and passive aggression, which are common in cyberbullying content (Awan & Zempi, 2017). Emotionaware systems analyze polarity and contextual sentiment shifts, as identical phrases may differ in meaning depending on context.

Transformer-based models such as BERT improve sentiment-aware detection by capturing deep contextual relationships and subtle semantic

variations (Bastiaensens et al., 2014). However, cultural and multilingual differences remain a major challenge, as expressions and sarcasm may vary significantly across languages and societies (Muneer & Fati, 2020; Ullah et al., 2022).

Despite their advantages, multimodal systems face challenges such as high computational cost, feature fusion complexity, and limited datasets. In addition, their interpretability remains limited, which raises concerns regarding transparency in automated moderation systems (Hosseini et al., 2017; MohammedJany et al., 2023).

Overall, multimodal and sentiment-aware approaches are considered essential for next-generation cyberbullying detection systems due to their ability to capture both emotional and contextual dimensions of online communication (Zampieri et al., 2019). Figure 2 illustrates the overall framework of multimodal and sentiment-based cyberbullying detection, including input modalities, feature extraction, multimodal fusion, sentiment analysis, and final classification stages.

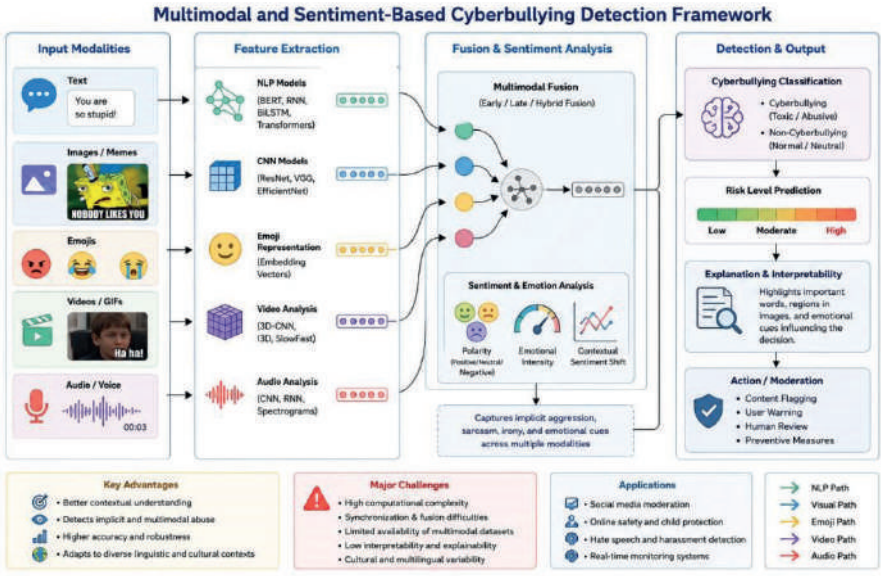


Figure 2. Multimodal and Sentiment-Based Cyberbullying Detection Framework

9. Evasion and Abuse of Cyberbullying Detection Systems

As cyberbullying detection systems become increasingly sophisticated, malicious users continuously develop new techniques to evade automated moderation mechanisms. These strategies are designed to bypass detection algorithms, manipulate classification outcomes, and reduce the effectiveness of AI-based content moderation systems. Consequently, evasion and abuse have emerged as significant challenges in the development of reliable cyberbullying detection technologies (Weimann & Masri, 2020; Weidinger et al., 2022).

One of the most common evasion strategies is text obfuscation. Users intentionally modify offensive words through misspellings, character substitutions, spacing alterations, and symbol insertion. For example, abusive terms may be disguised by replacing letters with numbers or special characters. Such modifications can make harmful content difficult for automated systems to recognize while remaining understandable to human readers. Obfuscation techniques are particularly effective against systems that rely heavily on keyword matching and surface-level textual features (Mozafari et al., 2019; Weimann & Masri, 2020).

Adversarial attacks represent a more sophisticated form of system manipulation. In adversarial attacks, malicious users deliberately construct inputs designed to deceive machine learning models. Small and seemingly insignificant changes to text can cause a model to misclassify harmful content as harmless. These attacks expose vulnerabilities in automated moderation systems and demonstrate the limitations of current classification approaches. Studies indicate that both traditional machine learning models and advanced AI systems may be susceptible to adversarial manipulation under certain conditions (Hosseini et al., 2017; Weidinger et al., 2022).

Meme-based attacks constitute another growing challenge. Online harassment is increasingly communicated through memes, edited images, screenshots, GIFs, and other visual formats. In many cases, the harmful message emerges from the interaction between visual content and textual context rather than from text alone. As a result, conventional text-based cyberbullying detection systems often fail to identify such forms of abuse. The growing popularity of visual communication has therefore increased the need for multimodal moderation approaches capable of analyzing both visual and textual information simultaneously (Yin et al., 2021; Wang et al., 2022).

Poisoning attacks target the training process rather than the deployment phase of AI systems. In these attacks, adversaries intentionally introduce misleading or malicious samples into training datasets. The objective is to

corrupt the learning process, reduce model accuracy, or create systematic blind spots that allow certain forms of cyberbullying to go undetected. Poisoning attacks are particularly concerning for systems that continuously learn from user-generated content because compromised training data may negatively affect future model performance (Selbst et al., 2019; Weidinger et al., 2022).

Another frequently observed strategy involves the use of coded language and evolving slang. Online communities often develop alternative vocabularies, abbreviations, and context-specific expressions that carry harmful meanings while avoiding detection by moderation systems. Because these linguistic patterns evolve rapidly, maintaining effective cyberbullying detection requires continuous monitoring and adaptation (Weimann & Masri, 2020; Jahan & Oussalah, 2023).

To counter these threats, researchers have proposed several defensive strategies. Adversarial training, data validation procedures, robust model architectures, and multimodal content analysis have been identified as promising approaches for improving system resilience. In addition, explainable AI techniques can help moderators better understand classification decisions and identify potential weaknesses exploited by malicious users (Hosseini et al., 2017; Wang et al., 2022).

Despite ongoing advances in AI-based moderation, the dynamic nature of online communication ensures that evasion techniques will continue to evolve. Future research should focus on developing adaptive, robust, and transparent detection systems capable of responding to emerging forms of cyberbullying while maintaining fairness, accountability, and user trust (Floridi & Cowls, 2019; Diaz-Garcia & Carvalho, 2024).

Table 5. Common Evasion and Abuse Techniques in Cyberbullying Detection Systems

Technique	Description	Potential Impact
Text Obfuscation	Intentional misspellings, symbols, and character substitutions	Bypasses keyword-based detection
Adversarial Attacks	Carefully crafted inputs designed to deceive AI models	Causes misclassification of harmful content
Meme-Based Attacks	Harmful content embedded in images, memes, and visual media	Evades text-only moderation systems
Poisoning Attacks	Injection of malicious samples into training datasets	Reduces model accuracy and reliability
Coded Language	Alternative vocabulary and evolving slang	Creates semantic ambiguity for detection systems
Multimodal Manipulation	Combining text and visual content to conceal harmful intent	Increases detection complexity
Adversarial Training	Training models using adversarial examples	Improves robustness against attacks
Adaptive Moderation Systems	Continuous updating of detection mechanisms	Enhances resilience against emerging threats

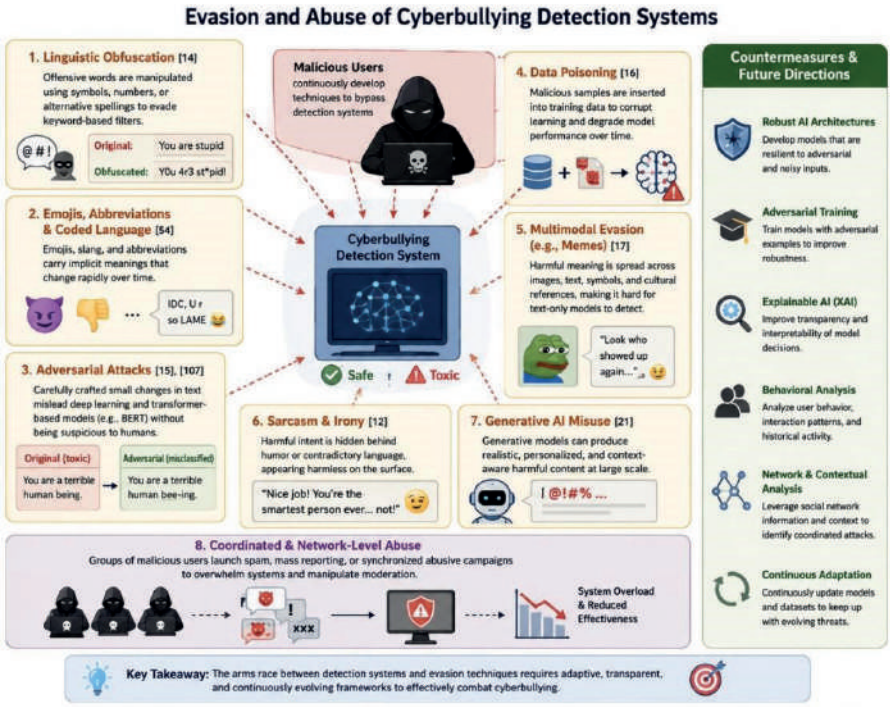


Figure 3. Overview of evasion techniques and abuse strategies in cyberbullying detection systems along with corresponding AI-based countermeasures.

10. Conclusion

The rapid expansion of social media platforms and digital communication technologies has transformed cyberbullying into a major social, psychological, and technological challenge. Unlike traditional forms of bullying, cyberbullying operates in highly dynamic and anonymous digital environments, where harmful content can spread rapidly and persist over long periods of time (AlAjlan & Ykhlef, 2018; Al-Dabet et al., 2023; Vidgen & Yasseri, 2020). These characteristics increase both the complexity of detection and the severity of its psychological and societal impacts, particularly among adolescents and vulnerable online communities (Aldreabi, 2024; Alabdulwahab et al., 2023).

This study reviewed the evolution of artificial intelligence-based cyberbullying detection and prevention systems, with a particular focus on machine learning, deep learning, and multimodal analytical approaches. Traditional machine learning methods such as Support Vector Machines and Random Forest algorithms have demonstrated useful performance in basic text classification tasks; however, their dependence on manual feature engineering and limited contextual understanding reduce their effectiveness in real-world social media environments (Aqeel & Kamble, 2022; LeCun et al., 2015).

Deep learning architectures, including CNN, RNN, LSTM, and hybrid CNN–LSTM models, have significantly improved cyberbullying detection performance by enabling automatic feature extraction and contextual learning from large-scale datasets (Meta Transparency Center, 2024; Aldreabi & Blackburn, 2023; SJ & Cho, 2020). Furthermore, transformer-based architectures such as BERT have introduced substantial advances in semantic understanding, allowing AI systems to better detect sarcasm, hate speech, implicit aggression, and context-dependent harmful language (Devlin et al., 2019; Mozafari et al., 2019; Mozafari et al., 2019).

The study also highlighted the growing importance of multimodal systems that integrate textual, visual, and sentiment-based analysis. Such systems provide more comprehensive understanding of online interactions and improve the detection of meme-based, symbolic, and visually embedded harmful content (Wang et al., 2022; Weimann & Masri, 2020; Zampieri et al., 2019). However, challenges such as adversarial attacks, linguistic obfuscation, multilingual variability, dataset imbalance, and algorithmic bias continue to limit the robustness and fairness of current detection systems (Biggio et al., 2012; Papernot et al., 2016; Selbst et al., 2019; Zhou et al., 2022).

In addition to technical challenges, ethical considerations remain central to the development of AIbased moderation systems. Issues related to user

privacy, transparency, freedom of expression, and explainability require careful attention in order to maintain user trust and prevent harmful consequences caused by incorrect moderation decisions (Hosseini et al., 2017; Weidinger et al., 2022; Selbst et al., 2019). Therefore, fully automated systems alone are unlikely to provide a complete solution to cyberbullying mitigation.

The findings of this study suggest that the future of cyberbullying prevention will increasingly depend on interdisciplinary and human-centered approaches that combine advanced AI technologies with ethical governance, digital literacy, regulatory frameworks, and human moderation mechanisms (Troop-Gordon et al., 2019; Aldreabi & Blackburn, 2023). Future systems are expected to become more adaptive, predictive, explainable, and multimodal, enabling earlier detection of harmful behavior and more accurate understanding of complex online interactions (Wang et al., 2021; Diaz-Garcia & Carvalho, 2024).

Ultimately, effective cyberbullying mitigation requires balancing technological innovation with ethical responsibility. Without integrating fairness, transparency, and human oversight into AI systems, even the most advanced detection technologies may remain insufficient for creating safe, inclusive, and sustainable digital environments.

References

Aldreabi, A. H. (2024). Artificial intelligence approaches for detecting Islamophobic hate speech on Reddit. Journal of Information Security and Applications, 79, 103650.

Aldreabi, A. H., & Blackburn, J. (2023). Explainable AI moderation systems for harmful online speech. IEEE Transactions on Artificial Intelligence, 4(6), 1238–1250.

Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep learning algorithm for cyberbullying detection. International

Journal of Advanced Computer Science and Applications, 9(9), 602–608. https://doi.org/10.14569/IJACSA.2018.090927

Alabdulwahab, S., Al-Khalifa, H., & Alageel, A. (2023). Artificial intelligence based prediction systems for online harmful behavior detection. Applied Sciences, 13(7), 4190.

Al-Dabet, S., Tedmori, S., & Al-Rawashdeh, M. (2023). Ethical considerations in AI-driven cyberbullying detection systems. IEEE Access, 11, 44120–44137.

Al-Hashedi, A., Altrjman, C., & Alshdaifat, E. (2022). A survey of machine learning and deep learning techniques for cyberbullying detection. Computers, Materials & Continua, 72(1), 561–587.

Aljalaoud, A., Alotaibi, N., & Althnian, A. (2022). Arabic cyberbullying detection using deep learning techniques. PeerJ Computer Science, 8, e896.

Aqeel, M., & Kamble, R. (2022). Hybrid machine learning framework for cyberbullying detection in social media. Expert Systems with Applications, 197, 116674.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82– 115.

Arshad-Ayaz, A., Naseem, M. A., & Khalid, S. (2020). Digital ethics and cyber citizenship in social media environments. Education and Information Technologies, 25(5), 3565–3583.

Ashraf, M., Ullah, A., & Khan, S. (2023). YouTube-based religious hate speech dataset for machine learning applications. Data in Brief, 49, 109319.

Atoum, Y. (2021). Sentiment analysis techniques for toxic language and cyberbullying detection. Journal of Big Data, 8(1), 114.

Awan, I., & Zempi, I. (2017). Cyber hate crime and the online abuse of Muslims. Palgrave Macmillan.

Bastiaensens, S., Vandebosch, H., Poels, K., et al. (2014). Cyberbullying on social network sites: An experimental study into bystanders' behavioural intentions. Computers in Human Behavior, 31, 259–271.

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In Proceedings of ICML.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877– 1901.

Castro, S., Hazarika, D., Pérez-Rosas, V., & Zimmermann, R. (2019). Towards multimodal sarcasm detection. In ACL 2019.

Dadvar, M., & Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models. arXiv preprint arXiv:1812.08046.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers. In NAACL-HLT 2019.

Diaz-Garcia, J. A., & Carvalho, A. (2024). Large language models for toxic content detection. Artificial Intelligence Review, 57(2), 1–29.

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for NLP. In ACL 2018.

Erliyani, N. (2021). Aggressive communication behaviors and empathy loss. International Journal of Social Psychology, 36(4), 518–531.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In WACV.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hinduja, S., & Patchin, J. W. (2019). Connecting adolescent suicide to bullying. Journal of School Violence, 18(3), 333–346.

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

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google Perspective API. arXiv preprint arXiv:1702.08138.

Hussain, Z., Griffiths, M. D., & Sheffield, D. (2018). Problematic smartphone use and cyberbullying. Computers in Human Behavior, 87, 269–276.

Jahan, I., & Oussalah, M. (2023). Context-aware NLP systems for online abuse prevention. KnowledgeBased Systems, 276, 110748.

Khalid, S. (2020). Cyberbullying and digital ethics in modern societies. Journal of Digital Society Studies, 12(3), 101–118.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

McMahan, B., Moore, E., Ramage, D., et al. (2017). Communication-efficient learning of deep networks.

In AISTATS 2017.

Mohammad, S. M. (2016). Sentiment analysis: Detecting emotions from text. In Emotion Measurement.

MohammedJany, K., Rahman, M., & Islam, T. (2023). Multimodal deep learning architectures for cyberbullying detection. Neural Computing and Applications, 35(19), 14125–14148.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection. In Complex Networks and Their Applications VIII.

Muneer, A., & Fati, S. M. (2020). Machine learning techniques for cyberbullying detection on Twitter.

Future Internet, 12(11), 187.

OpenAI. (2024). AI-generated content and provenance challenges. https://openai.com/research

Papernot, N., McDaniel, P., Swami, A., & Harang, R. (2016). Adversarial input sequences for RNNs. In MILCOM 2016.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Word representations. In EMNLP 2014.

Selbst, A. D., Boyd, D., Friedler, S. A., et al. (2019). Fairness and abstraction in AI systems. In FAT 2019.

SJ, T., & Cho, S. B. (2020). CNN-LRCN for cyberbullying detection. Sensors, 20(16), 4658.

Troop-Gordon, W., Gerardy, H., & Ladd, G. (2019). Cyber victimization and emotional well-being. Journal of Youth and Adolescence, 48(9), 1783–1796.

Ullah, I., Khan, A., & Lee, S. (2022). Multilingual abusive language detection. Applied Sciences, 12(14), 7142.

Van Hee, C., Jacobs, G., Emmery, C., et al. (2018). Automatic detection of cyberbullying. PLOS ONE, 13(10), e0203794.

Vidgen, B., & Yasseri, T. (2020). Detecting Islamophobic hate speech. Journal of Information Technology & Politics, 17(1), 66–78.

Wang, X., Zhao, Y., & Wang, H. (2022). Multimodal sentiment analysis. Information Processing & Management, 59(5), 103025.

Weidinger, L., Mellor, J., Rauh, M., et al. (2022). Ethical risks of language models. arXiv preprint arXiv:2112.04348.

Weimann, G., & Masri, N. (2020). Spreading hate on TikTok. Studies in Conflict & Terrorism, 46(5), 752– 765.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks at scale. In WWW 2017.

Xu, J., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces. In NAACL-HLT 2012.

Yadav, S., Ekbal, A., Saha, S., et al. (2021). Feature-assisted Bi-LSTM model. Knowledge-Based Systems, 213, 106704.

Yin, W., Zubiaga, A., & Wang, B. (2021). Multimodal abusive meme detection survey. ACM Computing Surveys, 54(7), 1–36.

Zampieri, M., Malmasi, S., Nakov, P., et al. (2019). Offensive language detection. In NAACL-HLT 2019.

Ziems, C., Held, W., Shaikh, O., et al. (2023). Large language models in computational social science. Computational Linguistics, 49(3), 1–39.

Chapter 2

Comparative Analysis of Innovative Thinking and Artificial Intelligence For Systematic Creativity

Kenan Peker[1]

Gökhan Önder Ergüven[2]

Abstract

In today's competitive environment, businesses struggle to achieve sustainable competitive advantage with traditional approaches that focus solely on reducing costs and increasing operational efficiency. In this context, this study examines innovation from a blue ocean strategy perspective. Innovation is considered a tool that enables businesses to grow through a blue ocean strategy, rather than a red ocean strategy. The study emphasizes the decisive role of innovation in enabling businesses to produce high value-added outputs. Within this framework, a case study is conducted on acquiring innovation skills, applying innovation theory to this purpose. The study demonstrates how innovation skills can be developed through theoretical applications. To better understand the abstract concepts of innovation, a case study using the sand-bicycle metaphor, developed within the framework of the SCAMPER approach, a thinking technique, is employed. Through this metaphor, it is shown that value originates not from physical inputs but from the transformation process achieved through knowledge, technology, and skills. In conclusion, the study suggests that businesses should build their competitive strategies not only on increasing efficiency in existing markets but also on innovation-focused approaches aimed at creating new value areas. In this respect, the study offers a conceptual and theoretical contribution to the innovation literature.

1 Munzur University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Production Management and Marketing Division, Tunceli, Turkey kpeker@munzur.edu.tr, ORCID:0000-0002-1573-2998

2 Munzur University, Faculty of Economics and Administrative Sciences, Department of Political Science and Public Administration, Division of Urbanization and Environmental Problems, 62000 Tunceli, Türkiye, gokhanondererguven@gmail.com, ORCID:0000-0003-1573-080X

https://doi.org/10.58830/ozgur.pub1351.c5535

1. Introduction

Artificial Intelligence (AI) has been very beneficial for getting info and knowledge as steps of science and technology production. In these movements, AI gathers data, and analyzing data with contacts and communication of data sources. It is not at the level of creative thinking yet. This study provides a comparative analysis between AI and a human for creativity and innovation as specific case though both lenses to evaluate their efficiency, depth of insight, and practical applicability. Because the sustainability of life within the socio-economic balance of the globalized world is under serious threat between AI and creative thinking. With globalization and digitalization, the world has become visible and accessible as if it were a village; consequently, value judgments that prioritize the social nature of human beings in competition have gradually been disappearing (Akbas et al., 2016). That's why today, under the prevailing "the strong wins" mentality, rights violations have increased within the framework of the red ocean strategy. Due to this ruthless competition, innovation has become an important subject of public policy in its theoretical, practical, experimental development, and social experimentation stages, and it is particularly regarded as the key to competitiveness (Gokce, 2015).

Innovation is a concept capable of opening every door and, by replacing the red ocean strategy with the blue ocean strategy, enabling access to resources sufficient for everyone. When the key to competition shifts from imitation, provocation, and aggressive confrontation toward genuine stages of innovation, a sustainable world for humanity will become possible (Kim & Mauborgne, 2015).

As is well known, individuals with weak character and insatiable desires often resort to deception and cunning. When the fundamental purposes of human activity to earn a living, ensuring the sustainability of life, and being beneficial (sustainability) are considered (profit, service to society, and sustainability), balance can be achieved if sustainability becomes the ultimate horizon. In other words, when differences are perceived as rich, a world in which culture, cultural diversity, knowledge, science, technology, innovation, development, and sustainability coexist will enable all living beings to live together in balance.

The mindset that frequently uses values to protect personal interests and transforms them into rent-seeking behavior gradually gives way to an approach that listens, speaks, reads, writes, imagines, discovers, and designs-produces. This state of mind forms the very foundation of the blue ocean strategy, which stands in contrast to the bloody competition inherent in the red ocean strategy. Innovative thinking has been very important for future of innovation.

Instead of focusing solely on earning a living and sustaining life, individuals direct their attention toward the broader objective of sustainability being beneficial. Through thoughts nourished by reason, they move toward knowledge-based development. Rather than undermining others in competition, they produce commercially viable outputs such as original designs, formulas, and products through new and innovative ideas (Stiglitz et al., 2009).

Such an understanding respects beliefs and nations and preserves the environmental balance in which plants and animals sustain their lives by organizing cultural activities enriched by the contributions of different cultures. In this way, goodwill and cultural diversity contribute to the construction of a world in which love, peace, and justice prevail for everyone and everything. Otherwise, unchecked ambition where individuals consume only what they themselves produce initially manifests itself in inconsistency and broken promises and ultimately leads to a situation in which one reaps what one sows.

They are pioneering leaders who stand out in communities and society with their humanity and good morals, wishing to benefit humanity; free in thought, original in ideas, and who will bring forth what is good for everything (Northouse, 2021). They are those who embark on a journey of knowledge, seeking to progress through stages that can be summarized as Faith, Islam, Sincerity, Mind, Thinking, Logic, Intelligence, Judgment, Knowledge, Wisdom, and Insight = Perfect Human. They aim to build and revive a blue ocean instead of competition, advancing with new ideas and innovations at every stage. They are not those who use national and spiritual values for personal gain, but those who love their country most with these sensitivities and work hard as a demonstration of this love. Rather than undermining those they perceive as competitors by putting forward various arguments against them while striving to secure the best for themselves, humanity should focus on sowing seeds and producing for the benefit of all living beings. The study proposes that numerous real-life events be examined in a similar manner to create "original designs based on readings of nature and history." It is predicted that by analyzing stories, tales, events, cases, etc., and applying innovation theories to foster innovation skills, the study will, over time, accelerate the creation and revitalization of blue-oceans through innovative entrepreneurship, replacing the red-ocean competitive approach with a blue-ocean strategy.

Today, the widespread accessibility of information and the necessity of coexistence among different cultures are causing individuals to face rapid and multifaceted changes in technological, social, cultural, and economic fields. In this world dominated by constant change and where the learning

process continues uninterrupted, individuals need to possess 21st-century skills to succeed (Bani-Hamad & Abdullah, 2019). These skills are defined as competencies necessary for problem-solving, critically approaching social issues, and achieving success in both professional and social life (Atalay & Boyacı, 2019).

High-tech industrial development zones are defined as special areas that bring together innovative resources to support the development of high-technology sectors, based on policy support and environmental advantages. Globally, high-tech zones have become a preferred development model for many countries and regions in terms of establishing innovation centers, nurturing innovative businesses, building innovation ecosystems, improving innovation performance, and guiding innovation-oriented economic development (Ulutas, 2020).

Learning and innovation skills refer to the mental processes necessary for individuals to adapt to and thrive in the modern work environment (OECD, 2019). While in the past, the storage and retrieval of information for future use was important, the proliferation of tools and technologies that directly guide individuals to readily available information has made learning and innovation skills even more critical (Kirschner and De Bruyckere, 2017). Integrating these skills into education systems is considered essential for effective development in the 21st century. In this context, stakeholders in education are making intensive efforts to ensure that students acquire learning and innovation skills through educational institutions.

Large-scale businesses and innovative and entrepreneurial businesses can operate in the same sector or in different sectors. The relationship between these two actors is too complex to be explained solely within the framework of "capital supply" and "capital demand." In the innovation and development process, the interaction between large businesses and entrepreneurial businesses does not exhibit a zero-sum structure; on the contrary, it is based on mutual value creation (Rossi et al., 2022).

Moore argued that businesses must continuously meet customer needs not only through competitive and collaborative dynamics, but also through innovation. In line with this approach, businesses have moved beyond being isolated industry actors and become part of a broader industrial ecosystem. Through analyses of companies such as Apple, IBM, Ford, and Walmart, Moore demonstrated that core businesses develop unique business ecosystems by creating service, technology, value networks, and generate economic value through these structures (Moore, 1999).

A strong symbiotic relationship exists between large corporations and startups. Examining this interactive symbiosis model between these actors and analyzing its evolution over time offers significant practical contributions and enriches the existing theory of innovation ecosystems (Durusoy, 2024).

Today, organizations are striving to maintain their presence in the global market due to challenges created by factors such as globalization, intense competition, and technological advancements. In this process, organizations are moving away from approaches focused solely on increasing efficiency and differentiating their products or services, and are turning to inimitable resources, especially human capital. Employees are considered the most important resource and asset for every organization; it is stated that organizations that manage their human capital effectively and efficiently are more likely to achieve their goals and ensure sustainable performance (Nafei, 2015).

Organizations today face problems related to talent limitations rather than capital shortages (Kehinde, 2012). Literature indicates that talented employees constitute only 3–5% of the total workforce in an organization (Berger & Berger, 2004). However, talent is considered a fundamental success factor in improving and sustaining organizational performance. The concept of talent encompasses an individual's skills, experience, knowledge, intelligence, and qualifications, as well as their capacity for learning and development (Nafei, 2015).

These regional collaboration networks are organized by public institutions, businesses, universities, research centers, and financial institutions; and share common characteristics such as internal cooperation, embeddedness, openness, stability, and dependence on the environment (McPhilips, 2020). Such regional knowledge networks play a decisive role in the formation of regional economic competitiveness, development potential, and regional advantages by enabling the efficient allocation of heterogeneous knowledge sources and the production of new knowledge.

Today's businesses must make strategic decisions with a more rational and long-term perspective due to increasing competitive pressure, limited resources, and rapidly changing consumer demands. In this context, increased efficiency is considered not only an operational performance indicator but also a strategic competitive tool. However, strategies based on direct competition with rivals in existing markets make it difficult for businesses to achieve sustainable competitive advantage in the long term. At this point, the blue ocean strategy gains importance as an innovative approach that redefines the boundaries of competition. For the blue ocean strategy gains, creative

thinking is necessary and AI have to be developed by innovative thing where some techniques necessary such as SCAMPER, and TRIZ.

2. SCAMPER technique for Creative Ideas

The SCAMPER technique representing Substitute, Combine, Adapt, Modify, Magnify, Minimize, Put to other use, Eliminate, Reverse, and Rearrange is a creative thinking strategy designed to move designers beyond conventional logic and stimulate a wide spectrum of innovative ideas (Boonpracha, 2023). Developed by the American psychologist Robert F. Eberle, SCAMPER functions as an intuitive and user-friendly instructional approach for creativity, widely applied in product development and project enhancement processes (Tharwa & Farid, 2019).

This method offers practical tools that support idea generation, helping individuals overcome the psychological challenge often associated with confronting a "blank page" and enabling a transition toward creative thinking. Through a series of guiding prompts, SCAMPER encourages multidimensional and expansive thinking, thereby strengthening both the range and depth of cognitive engagement. Owing to its clear structure and comprehensive framework, the technique can be easily learned and effectively applied in diverse contexts.

3. AI Driven Innovation Ecosystems

Artificial Intelligence (AI) has become one of the most influential technological developments of the contemporary era, significantly transforming industries and altering competitive dynamics. However, the complexity and high costs associated with the development and implementation of AI technologies, along with uncertainties regarding value generation, indicate that organizations are unlikely to fully exploit AI's capabilities on their own. Consequently, this situation has encouraged the emergence of innovative ecosystems in which various organizations collaborate by combining complementary capabilities, pooling resources, and fostering collective advancement. Furthermore, as a general-purpose technology, AI has the capacity to reshape and disrupt established innovation ecosystems, including fields such as robotics, pharmaceutical development, and bioengineering.

Moore was the first academic to define the conceptual content of the innovation ecosystem, arguing that the interdependence among different actors in the system reflects ecological characteristics. According to Moore, the group organizational ecology created by innovative businesses in accordance with specific rules and orientations constitutes the innovation ecosystem.

The introduction of this concept triggered widespread research in academic circles. Most researchers state that the innovation ecosystem consists of interconnected and interdependent network participants (main businesses, customers, suppliers, complementary innovators, and regulatory bodies), and that these participants also demonstrate absolute dependence on the system environment (Gomes et al., 2018).

The emergence of the innovation ecosystem concept is based on a deepening understanding of the innovation system and the continuous development of innovation practices. Innovation ecosystem theory is a theoretical system that incorporates the fundamental approaches of ecological theory and evolutionary economics theory, and represents the latest stage in the deepened development of innovation systems theory. This theory highlights the dynamic growth characteristics of the innovation system and emphasizes its self-organizing nature (Zeng et al., 2013).

Innovation ecosystems, like biological systems, have evolved from a structure composed of initially randomly selected elements to an ordered and structured community; however, they still retain their essence as an innovation system (Iansiti and Levien, 2004). Feng et al. state that the innovation ecosystem is structurally composed of innovation ecological communities, and these communities consist of populations with different sources of innovation. The innovation ecosystem is defined by features such as nestedness, multi-layeredness, and multiplicity, and is considered a typical complex network system that transcends physical boundaries (Feng and Yang, 2020).

4. Machine Learning Innovation Ecosystems

At present, innovation management organized by human actors remains central to firms and to their ability to renew themselves through exploratory activities. Nevertheless, Artificial Intelligence (AI) can provide forms of support that extend beyond human capabilities (Wamba et al., 2017). Both scholars and industry practitioners have suggested that AI is likely to exert a considerable influence on organizational innovation processes in the future (Bughin et al., 2018). This perspective is reinforced by the rapid progress in AI and machine learning technologies, which signals the potential for significant and far-reaching transformations (Varian, 2018). Despite these developments, the current understanding of the limitations of AI within innovation contexts remains relatively limited. Moreover, applying AI and machine learning to creativity and innovation differs substantially from more established domains in which AI has already replaced conventional managerial functions (Chui et al., 2018).

5. Technological Innovation Efficiency

Technology innovation efficiency is generally defined as the ratio of the outputs obtained in the technology innovation process to the inputs transferred to the process. This ratio is considered an important indicator in evaluating whether resources are allocated effectively and in assessing the innovation capacity of businesses. In the literature, Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA) are among the most commonly used methods for measuring technology innovation efficiency (Gong et al., 2020).

6. Data Analytics for Talent Management

Data analytics for talent management plays a critical role in fostering a supportive and growth-oriented environment for employees, ensuring that appropriate resources and opportunities are directed toward the right individuals to achieve strategic objectives. The ability to effectively address future organizational challenges significantly influences the sustainability and success of various departments. Nonetheless, risk management efforts cannot succeed without competent personnel. This underscores the importance of talent management practices, which aim to align with the organization's overall strategy by attracting, motivating, developing, and retaining highly skilled and capable employees (Aina & Atan, 2020). Talent represents a pivotal factor in enhancing and sustaining organizational performance, encompassing not only skills, knowledge, and experience but also intelligence, personal competencies, and the capacity for continuous learning and development (Berger, 2004). Research investigating the influence of talent management on organizational outcomes operates on the premise that effective talent management contributes to improved performance by securing and retaining individuals with the requisite abilities (Armstrong & Taylor, 2014).

7. Big Data in Strategic Innovation

Big data refers to extremely large data sets generated and accumulated by organizations, which continue to expand rapidly over time. Due to their scale and complexity, these datasets are difficult to manage using conventional data-processing software. As a result, specialized big data technologies and tools have been developed to support organizations in collecting, processing, and analyzing large volumes of information. These tools enable firms to derive meaningful insights, address complex business challenges, enhance organizational agility and innovation, and support more effective decision-making processes (Alghamdi & Ahag, 2023). Essentially, the objective of big data is to extract intelligence from large datasets in order to identify

opportunities and convert them into competitive business advantages (McAfee et al., 2012).

However, the mere accumulation of large datasets without a clear objective may not provide long-term value (Müller, 2016). To produce meaningful outcomes, big data must be systematically organized and analyzed through appropriate analytical methods, tools, and techniques in order to uncover valuable insights and support effective data visualization (Mikalef et al., 2018). In this regard, big data analytics refers to the process of gathering, examining, and presenting insights derived from large datasets in ways that facilitate actionable knowledge, generate business value, and strengthen competitive advantage (Wamba et al., 2020). Successfully implementing this process requires adequate organizational resources, specialized skills, and well-developed capabilities to ensure effective execution.

8. Methodology

This methodology of the study contributes to innovation literature in several ways. First, it provides a **conceptual integration of productivity, talent management, innovation ecosystems, and Blue Ocean Strategy**, offering a holistic framework for understanding innovation processes.

Second, the study introduces the **sand–bicycle metaphor developed within the SCAMPER framework**, which serves as an analytical tool for explaining how value can emerge from the transformation of ordinary inputs through knowledge and creativity.

Third, the study contributes methodologically by demonstrating how **creative thinking techniques can be used as explanatory models in innovative research**.

Finally, the study highlights the strategic importance of **human capital and talent management** in innovation ecosystems and emphasizes that sustainable competitive advantage depends on knowledge-based transformation rather than physical resources alone.

10. Competition and Productivity in Businesses

In the context of businesses, productivity means minimizing costs, using resources effectively and rationally, and optimizing business processes. High levels of productivity not only provide businesses with an advantage in price competition but also strengthen other competitive elements such as quality, flexibility, and service level (Oren, 2016). While increased productivity has a profitability-enhancing effect in the short term, it plays a supportive role in the

innovation capacity and investment potential of businesses in the long term. In this context, productivity is of strategic importance in achieving sustainable competitive advantage. This requires the development of innovation skills and a blue ocean strategy.

The Blue Ocean Strategy was introduced to literature by Kim and Mauborgne (2005). The authors defined traditional competitive environments as "red oceans," arguing that intensified competition in these environments leads to decreased profitability. The Blue Ocean Strategy, however, aims to render competition irrelevant by creating new areas of demand and simultaneously reduce costs and increase customer value through value innovation. The Blue Ocean approach prevents businesses from focusing solely on existing competitors; it also encourages them to explore potential demand and redefine market boundaries. In this context, Blue Ocean strategies offer businesses the opportunity to gain a sustainable and long-term competitive advantage.

Productivity enhancement and the blue ocean strategy are considered two complementary strategic elements. Businesses with efficient processes can allocate more resources to innovation activities and thus increase their capacity to create a blue ocean. On the other hand, new business models developed within the framework of blue ocean strategies contribute to further increases in the productivity levels of businesses. This relationship is observed more clearly in sectors with intensive technological innovations; digital transformation processes provide businesses with both increased productivity and the opportunity to develop completely new value propositions (Yakar and Tasliyan, 2025).

The concept of value innovation forms the basis of the blue ocean strategy and explains the relationship between efficiency and innovation. Kim and Mauborgne (2015) showed that in successful blue ocean examples, businesses create new areas of benefit for the customer by eliminating unnecessary cost elements. This approach demonstrates that efficiency is not only a tool for cost reduction but also an element that supports strategic innovation.

In the literature focusing on technology and digital transformation, the role of increased productivity in creating a blue ocean is becoming more prominent. Brynjolfsson and McAfee (2014) state that digital technologies increase the productivity of businesses while simultaneously enabling the emergence of new business models. This reveals that when increased productivity is integrated with strategic innovation, it transforms into a competitive advantage.

In conclusion, the literature shows that productivity increases are necessary for competitive advantage but not sufficient on their own; sustainable value can

be created when supported by innovative approaches such as the blue ocean strategy. This study, building upon this literature, examines the complementary relationship between productivity and the blue ocean strategy within a holistic framework.

11. Productivity and Organizational Performance Through Talent

The concept of "talent" is considered the most valuable component of human capital and is treated as a strategic resource in modern management literature (Barkun et al., 2020). In literature, talent management is defined as the process of selecting, developing, motivating, and retaining employees, and it is frequently emphasized that these processes increase employee productivity and organizational performance. Various studies show that talent management practices have positive effects on employee performance and overall productivity. Integrating management into organizational strategies not only increases efficiency but also provides a sustainable competitive advantage.

The literature contains studies examining the relationship between talent management and productivity in various sectors (healthcare, manufacturing, services, etc.); however, most of these studies are Western-centric, and cultural differences have been addressed to a limited extent. While previous research has shown a significant relationship between talent management and organizational performance (Stahl et al., 2012), how talent management practices should be implemented to ensure sustainable organizational performance remains a controversial issue. Furthermore, a large portion of these studies have been conducted in the USA and Western Europe, where talent management is considered a mature practice. These countries have recognized the critical role of human capital in the development of organizations and nations; they possess an advanced and adaptable structure in terms of civilization, economy, and technological infrastructure, as well as the capacity to implement new techniques and practices.

Although many organizations in developing countries, particularly in the Middle East, have recently adopted talent management practices from developed countries, researchers recommend that organizations in these countries not simply copy these practices. Factors such as cultural differences, structural imbalances, religious and cultural conflicts, and underdeveloped financial markets can limit the effectiveness of the practice (Gandhok and Smith, 2014).

12. Analytical Model

The analytical model of this study is based on the premise that the perceived value of natural resources can be transformed through innovation-oriented thinking frameworks. In this context, the SCAMPER technique is adopted as an analytical tool to explore how a traditional raw material such as sand can be reconceptualized and utilized within modern industrial production systems. The model assumes that each component of the SCAMPER framework (Substitute, Combine, Adapt, Modify, Put to Another Use, Eliminate, and Reverse/Rearrange) represents a distinct mechanism for generating innovative perspectives on resource utilization.

Within this framework, sand is considered the primary input resource, while technological processes, knowledge-based capabilities, and human capital represent enabling factors that facilitate value transformation. The analytical model suggests that the interaction between these elements leads to the emergence of new industrial applications and higher value-added products. In particular, the transformation of sand into materials such as silicon and fiberglass illustrates how innovation processes can convert traditional raw materials into strategic technological inputs.

Accordingly, the analytical model conceptualizes value creation as a multi-stage process in which raw materials are reinterpreted through innovation mechanisms and integrated into advanced manufacturing systems. This perspective highlights that industrial value is not determined solely by the physical characteristics of resources but by the knowledge, creativity, and technological capabilities that enable their transformation.

13. Conceptual study: The Sand–Bicycle Metaphor and SCAMPER Analysis

To illustrate the development of innovation skills, this study introduces a metaphorical case based on the SCAMPER creative thinking technique.

To illustrate this point with a well-known anonymous story, let's talk about the story of the Mexican biker. The story describes a man repeatedly crossing a border with a sack filled with sand on his bicycle. Border officials repeatedly inspect the sack but fail to realize that the object being smuggled is the bicycle itself.

This example highlights a common cognitive bias: focusing excessively on complex details while overlooking the most obvious element.

Within the SCAMPER framework, the metaphor demonstrates several innovation principles:

A Mexican man was entering the U.S. on a bicycle, with a huge sack on the rear rack. The border officer became suspicious and said, "Open the sack." The Mexican opened it. Sand. The next day, the same Mexican showed up at the border whistling, again with a sack on the back of his bicycle. The officer grew suspicious again. "Open it." Again, sand. The following day, the same Mexican pedaled up to the border crossing, once more with a sack on the rack. This time, not just customs officers but narcotics officers had him open it. But it was useless out came sand yet again. They were going crazy. Five days, two weeks, a month is always the same scene. The Mexican casually arrives; they scan him with thermal cameras, run tests, let the dogs sniff around, track him with satellite images it's all in vain. The sack contains sand.

Years pass… One day, the border officer who had nearly lost his mind over this run into the Mexican at a bar. He says, "It's eating me up inside. What's done is done. At this point I can't do anything to you. For heaven's sake, tell me what you were smuggling in that sack?"

"Bicycles," he says!

When we asked Artificial Intelligence, it gave the following response:

Artificial Intelligence's Comment on the Story:

The clever twist in this joke comes from focusing attention on the content while overlooking the vehicle itself.

The item being traded here: the bicycle.

The Logic of the Situation:

While the officers were busy each time with the "sand" inside the sack, they failed to notice that the Mexican was bringing a new bicycle across the border every day. The police assumed so (Anonymus)

This story perfectly illustrates how the most obvious thing can become invisible when we focus too much on complicated details (OpenAI, 2026)

A Mexican man attempts to enter the United States by bicycle, carrying a large bag on the saddle. When the police open the bag, they find only sand inside. Similarly, on subsequent entries, the bag is again filled only with sand. Years later, even with advanced technology and trained dogs, border security officials examine the bag, but the result remains the same: only sand. Finally, the police ask the Mexican man, "What was in the bag?" and his answer is, "A bicycle." This story shows that what is truly valuable is not the bicycle itself, but the "innovative idea" carried in the bag. In other words, the Mexican man went to the US to gain innovation, and the bag symbolizes that innovation.

The sand in the Mexican's sandbag may not seem like a direct raw material in bicycle production, thanks to modern manufacturing technologies, it plays a critical role in the production process of almost every part of a bicycle. What creates value is not the visible raw material (sand), but its transformation with a different function and technology. Instead of sand being a seemingly worthless and ordinary raw material, it undergoes a perceptual and functional substitution as a fundamental input for high value-added innovation. Therefore, in the question "What is the real value?", knowledge, technology, and skills are substituted for the physical product. This directly corresponds to the "Substitute" heading in SCAMPER. The sandbag example, while not standalone, also includes the following SCAMPER steps at a secondary level.

Reframe/Reverse (Reverse/Rearrange: Related to "Reverse/Rearrange" in SCAMPER. While the police focus on the bag, the real value is hidden in the bike.)

This is a reversal of the question "Where is value to be found?" and a shift from concrete input to abstract ability.

Put to Another Use: Sand is normally a low-value raw material. In the presented example, it has been used for multiple high-tech purposes such as microchips, composite materials, glass fiber, electronic components, and surface technology. This clearly falls under the scope of Put to Another Use.

Combine: In the example: The raw material, sand, is integrated under a single metaphor by considering Technology, Talent management, and the Innovation ecosystem together. This supports the Combine dimension in SCAMPER.

Here's how sand transformed into bicycle parts:

Metal Part Casting (Molding): The bicycle frame (whether aluminum or steel), pedals, brake calipers, and gear systems are usually produced using the casting method.

Sand Molding: Molds are prepared from sand to shape the metal. Molten metal is poured into these sand cavities. In other words, without sand, it would be much more difficult to obtain those complex metal forms.

Fiberglass and Carbon Fiber Parts

The lightweight components found in high-end bicycles are based on silica (sand).

Rims and Frame : Fiberglass is obtained from sand melted at high temperatures. Some composite bicycle parts use these sand-derived fibers to increase durability.

Electronic Components (Gears and Indicators): If your bike is electric or has a wireless shifting system (such as SRAM eTap, Shimano Di2), you're dealing with the most technologically advanced version of sand riding.

Microchips: The silicon in the sand is the main component of processors and sensors. Bicycle computers and electronic gear control units operate using this technology derived from sand.

Surface Treatments (Sandblasting): After parts are manufactured, the paint needs to adhere well to the metal or achieve a matte finish.

Sandblasting: Before painting, the bicycle frame is cleaned and smoothed using high-pressure sandblasting.

In summary, the sand tracks on the bicycle can be presented in Table 1.

Table 1. The role of sand in bicycle manufacturing.

Industry / Part Group	**Functional Role of Sand**	**Process Type**	**Material Used**	**Industrial Purpose**
Tooling and Surface Preparation	Shaping and cleaning	Sandblasting process	Sandblasting molds/abrasive sand	Surface finishing and preparation
Automotive Components (Gears&Brakes)	Mold formation in casting	Metal casting	Silica sand	Production of metal components
Electronics Industry (Displays)	Semiconductor production	Circuit and chip manufacturing	Silicon	Production of electronic devices
Composite Materials	Fiber reinforcement production	Composite manufacturing	Fiberglass	Manufacturing durable composite parts

The table illustrates the functional roles assumed by sand and sand-derived materials across different industries. The analysis reveals that sand is not only utilized in traditional casting processes but also serves as a critical input in advanced manufacturing sectors such as electronics production and composite material technologies. In particular, silica sand is widely used in the metal casting industry for mold production, whereas silicon derivatives constitute the fundamental raw material for semiconductor manufacturing in the electronics sector. Similarly, the production of fiberglass enhances the

mechanical strength of composite materials, thereby playing a significant role in modern manufacturing technologies. Overall, these findings indicate that sand-based materials represent a versatile and strategically important resource across various industrial processes.

14. Framework

This research framework explains the role of sand-based raw materials in modern industrial production systems. The model assumes that different types of sand-derived materials, such as silica sand, silicon, and fiberglass, function as key input resources in various industrial processes. These raw materials undergo several processing mechanisms including sandblasting, metal casting, semiconductor fabrication, and composite manufacturing. Through these processes, sand-based materials contribute to multiple industrial application areas such as the automotive, electronics, and composite materials industries. Ultimately, these applications influence industrial outputs, including manufacturing efficiency, product durability, and technological performance. The framework therefore conceptualizes sand-based materials as strategic resources that enable diverse manufacturing technologies and industrial production systems.

This example metaphorically illustrates the strategic role of talent management and innovation on organizational performance. When managed correctly, talent provides businesses with productivity, innovation, and a sustainable competitive advantage.

The multifaceted relationship between talent management, innovation, and productivity is often discussed through abstract concepts, making it difficult to concretize at the application level. Therefore, to present this relationship in a more understandable and analytical way, a metaphorical example is used in this study, employing the SCAMPER approach, a creative thinking technique. The SCAMPER technique offers a systematic framework that allows for the rethinking of existing elements in different ways and serves as an explanatory tool in explaining how innovation emerges. In this context, the sand-bicycle metaphor presented below aims to demonstrate that even low-value-added inputs can be transformed into high-value-added outputs when talent and knowledge are managed correctly. Table 2 shows the SCAMPER analysis of the sand-bicycle metaphor.

Table 2. SCAMPER Analysis of the Sand-Bicycle Metaphor

SCAMPER Dimension	Conceptual Transformation	Innovation Mechanism	Industrial / Strategic Outcome
Substitute	Sand is reconsidered not as a low-value raw material but as a strategic technological input.	Cognitive reframing of resource value	Recognition of knowledge-driven value creation
Combine	Sand, technology, and talent management are integrated.	Innovation ecosystem approach	Synergy between natural resources and human capital
Adapt	Sand is adapted to different production technologies.	Process innovation	Integration of traditional resources into advanced manufacturing
Modify	Sand is transformed into silicon, glass fiber, and other advanced materials.	Material transformation and technological upgrading	High value-added industrial products
Put to Another Use	Sand is used in semiconductors, composite materials, and surface technologies.	Functional diversification	Expansion of industrial application areas
Eliminate	The assumption that raw materials alone generate value is rejected.	Paradigm shift in production logic	Emphasis on knowledge, technology, and skills
Reverse / Rearrange	The perspective shifts from raw material to value-creation capability.	Strategic perspective change	Focus on innovation and capability-based value

The sand-bike metaphor, within the scope of the SCAMPER technique, is **largely based on the "Substitute"** approach. In this metaphor, the traditional assumption that value originates from raw materials is replaced by the idea that knowledge, skill, and innovative transformation capacity are the primary sources of value. In this respect, the example aligns with the fundamental assumptions of value innovation and blue ocean strategies.

However, the metaphor also includes the steps of **"Put to Another Use"** and **"Reverse** ." Demonstrating the uses of sand in different sectors and advanced technologies reveals how the relationship between productivity and innovation can be strengthened through the strategic repositioning of inputs. Seeking value not in the physical object itself, but in how it is transformed, supports the systems perspective emphasized in the innovation ecosystem literature.

In this context, the sand-bike metaphor is not only an illustrative narrative element but also an example of innovation based on analytical and theoretical foundations, compatible with the SCAMPER technique.

15. Conclusion and discussion

This study examined the relationship between productivity, innovation, and strategic competitiveness within the framework of Blue Ocean Strategy and innovation ecosystems. The findings suggest that organizations should not rely solely on operational efficiency but should also invest in innovative capabilities and talent development. By using the SCAMPER technique and the sand–bicycle metaphor, the study demonstrates that innovation emerges from the transformation of resources through knowledge, technology, and creativity.

The relationship between productivity increases and strategic innovation within a holistic framework, considering the intense competition, rapid technological transformation, and limited resources faced by today's businesses. Literature findings indicate that while productivity is a significant factor supporting profitability in the short term, it is insufficient on its own to create a sustainable competitive advantage. Therefore, this study emphasizes the necessity of considering productivity-focused approaches in conjunction with blue ocean strategies and value innovation. In this research, innovation ecosystems were evaluated as dynamic structures that enable knowledge sharing and shared value creation among businesses, universities, research centers, and public institutions. Within these ecosystems, human capital and talent management stand out as one of the key determinants of innovation capacity. Consistent with findings in the literature, talent management practices were found to improve employee performance, support organizational learning, and thus contribute to both increased productivity and the creation of innovative outputs.

One of the study's original contributions is the concretization of the abstract relationships between talent, productivity, and innovation through the sand-bike metaphor developed within the framework of the SCAMPER technique. This metaphor demonstrates that value stems not from physical inputs, but from how these inputs are transformed through knowledge, technology, and skills. The prominence of the Substitute approach in the SCAMPER analysis reveals that the traditional understanding of value, centered on raw materials, is being replaced by value creation based on knowledge and skills. This finding also aligns with the fundamental assumptions of the blue ocean strategy.

However, the fact that the sand-bike metaphor also includes the "Put to Another Use" and "Reverse" dimensions shows that innovation is not limited to new product development; it can also arise through the repositioning of existing resources in different ways. This highlights the importance of actors in innovation ecosystems re-evaluating their limited resources from a strategic perspective.

The findings of this study not only offer significant managerial implications for businesses but also highlight the need for organizations to focus solely on operational efficiency. In this context, it addresses talent management at a strategic level, replacing traditional competitive strategies. It develops approaches integrated with innovative ecosystems. Productivity increase should not be considered an end, but rather a tool enabling strategic innovation. In conclusion, this study demonstrates that the relationship between productivity, talent management, and innovation is not linear, but rather a multi-dimensional and dynamic ecosystem-based one. Productivity-focused approaches, supported by a blue ocean strategy, not only provide businesses with a competitive advantage in existing markets but also create new value areas, paving the way for sustainable growth. Future research testing this conceptual framework empirically in different sectors and countries will make significant contributions to literature.

Future research may extend this conceptual framework by conducting empirical studies across different industries and countries to test the relationship between talent management, productivity, and innovation performance. Applied research, experimental development, and social experiments are succents for future research about innovative thinking and innovative culture where blue ocean strategy impacts productivity without destroying competitors. Business worlds must move very fast to contact, communicate, collaborate, cooperate, cluster, create together for community development. Otherwise, some competitors will be destroyed due to red ocean strategy since AI uses becoming widespread in business worlds. As understood from the bicycle metaphor, AI is not going to replace the human brain. Humans need to use AI for knowledge and continue with creative thinking for innovation where blue economy occur without destroying competitors as well as environment.

References

Al Aina, R., & Atan, T. (2020). The Impact of Implementing Talent Management Practices on Sustainable Organizational Performance. Sustainability, 12(20), 8372

AMH Bani-Hamad et al. (2019). The effect of project-based learning to improve the 21st century skills among Emirati secondary students. International Journal of Academic Research in Business and Social Sciences.

Armstrong, M., & Taylor, S. (2014). Armstrong's Handbook of Human Resource. United Kingdom: Kogan Page

Anonymous 2026. AI tools such as gemini, and chatgpt for basic uses.

Akbas, Z., Babahanoglu, V., Cayli, S. (2016). The Effects of Capitalist Globalization on the Socio-Economic and Political Field in the Middle East: An Opportunity or a Threat for Stability and Development? Düzce University Journal of Social Sciences, 6(2), 86-108.

Alghamdi, O.A.; Agag, G. Boosting Innovation Performance through Big Data Analytics Powered by Artificial Intelligence Use: An Empirical Exploration of the Role of Strategic Agility and Market Turbulence. Sustainability 2023, 15, 14296.

Atalay, N., Boyacı, S (2019). Slowmation application in development of learning and innovation skills of students in science course. International Electronic Journal of Elementary Education.

Barkun, Y., Rollnik-Sadowska, E., & Glińska E. (2020). The Concept of 'Talent' in the Labor Management Perspective – The Bibliometric Analysis of Literature. International Journal of Industrial Engineering and Management, 11(2), 104–115.

Berger, L. A. (2004). The Talent Management Handbook: Creating Organisational Excellence by Identifying, Developing and Promoting Your Best People. New York: McGraw-Hill

Berger, L.A., Berger, DR (Eds.) (2004). The Talent Management Handbook: Creating Organizational Excellence by Identifying, Developing and Promoting Your Best People; McGraw-Hill: New York, NY, USA.

Boonpracha, J. SCAMPER for creativity of students' creative idea creation in product design Thinking Skills and Creativity, 48 (2023)

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N., Trench, M., 2017. Artificial intelligence: The next digital frontier? McKinsey Global Inst

Brynjolfsson, E. ve McAfee, A. (2014), The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York, NY: WW Norton & Company

Chui, M., Henke, N., Miremadi, M., 2018. Most of AI's business uses will be in two areas. Harv. Bus. Rev. 3–7

Durusoy, O. T. (2024). Industrial Symbiosis: An Environmental Approach. Ankara Hacı Bayram Veli University Journal of Faculty of Economics and Administrative Sciences, 26(2), 563-590.

Feng, W., Yang, S. (2020). High-tech industry agglomeration measurement and comparative research: An empirical analysis based on China's 2007–2017 data . J.Ind. Technol. Econ., 6, 154-160

Gandhok, T.; Smith, R. Rethinking (2014). Cross-Border Talent Management: The Emerging Markets Perspective. Asian Manag. Insights, 1, 18–25

Gokce, S.M. (2015). Innovation in Public Sector and its Applications in Türkiye. International Journal of Management and Social Research. 2(1): 28-35

Gomes, L. A., Facin, A. L., Salerno, M. S., & Ikenami, R. K. (2018). Unpacking the innovation ecosystem construct: Evolution, gaps and trends. Technological Forecasting and Social Change, 30-48

Gong. et al. (2020). Market acceptability assessment of electric vehicles based on an improved stochastic multicriteria acceptability analysis-evidential reasoning approach J. Clean. Prod.

Iansiti, M., Levien, R. (2004). Strategy as Ecology. Harv. Bus. Rev., 82, 68-81

Kehinde, J. (2012). Talent Management: Effect on Organization Performance. J.Manag. Pic., 4, 178–186

Kim, W. C., & Mauborgne, R. (2015). Blue ocean strategy: How to create uncontested market space and make the competition irrelevant (Expanded ed.). Harvard Business School Press.

Kirschner, P.A., & De Bruyckere, P. (2017). The myths of the digital native and the multitasker. Teaching and Teacher Education, 67, 135–142.

McAfee, A.; Brynjolfsson, E.; Davenport, T.H.; Patil, D.J.; Barton, D. (2012). Big data: The Management Revolution. Harv. Bus. Rev. 90, 61–67

McPhillips, M. (2020). Trouble in Paradise? Barriers to Open Innovation in Regional Clusters in the Era of the 4th Industrial Revolution. J. Open Innov. Technol. Mark. Complex., 6

Mikalef, P.; Pappas, I.; Krogstie, J.; Giannakos, M. Big data analytics capabilities: A systematic literature review and research agenda. Inf. Syst. E-Bus. Manag. 2018, 16, 547–578.

Moore, J.F. (1999). Predators and prey: A new ecology of competition. Harvard Business School Press

Müller, O., Junglas, I., Brocke, J.V., Debortoli, S. Utilizing big data analytics for information systems research: Challenges, promises and guidelines. Eur. J. Inf. Syst. 2016, 25, 289–302.

Nafei, W.A. (2015) Talent Management and Health Service Quality from the Employee Perspective: A Study on Teaching Hospitals in Egypt. Am. Int. J. Soc. Sci., 4, 91–110

Northouse, P. G. (2021). Leadership: Theory and practice (9th ed.). Sage Publications. OECD skills outlook 2019: Thriving in a digital world. OECD Publishing.

OECD 2019. https://www.oecd.org/content/dam/oecd/en/publications/reports/2019/05/oecd-skills-strategy-2019_g1g9ff20/9789264313835-en.pdf

Open AI, January, 15, 2026, https://chat.openai.com/chat

Oren, K. (2016). Time Management in Increasing Productivity and Performance in Businesses. HAK-İŞ International Journal of Labor and Society. 5(11): 186-203.

Rossi, M., Chouaibi, J., Graziano, D., Festa, G. (2022). Corporate venture capitalists as entrepreneurial knowledge accelerators in global innovation ecosystems. Journal of Business Research, 142 pp. 512-52.

Stahl, G.K.; Bjorkman, I.; Farndale, E.; Morris, SS; Paauwe, J.; Stiles, P.; Trevor, J.; Wright (2012). PM Six principles of effective global talent management. MIT Sloan Manag. Rev. 53, 24–32

Stiglitz, J.E., Sen, A., & Fitoussi, J.-P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress. OECD.

F. Tharwa, F. Farid. Using the SCAMPER model to develop translation skills for major students in the faculty of education. AWEJ for Translation & Literary Studies, 3 (2) (2019), pp. 91-113

Ulutas, F. (2020). The Impact of Technology Development Zones on the National Economy: An Evaluation Based on the Perceptions of ADU Technopark Company Managers. Master's Thesis, Department of Human Resources Management, Institute of Social Sciences, Aydın Adnan Menderes University. 2020

Varian, H., 2018. Artificial intelligence, economics, and industrial organization (No. 24839). In: NBER Working Paper, NBER Working Paper Series. Cambridge, MA

Wamba, S.F., Gunasekaran, A., Akter, S., Ren, S.J.-F., Dubey, R., Childe, S.J., 2017. Big data analytics and firm performance: effects of dynamic capabilities. J. Bus. Res. 70, 356–365

Wamba, S.F.; Dubey, R.; Gunasekaran, A.; Akter, S. The performance effects of big data analytics and supply chain ambidexterity: The moderating effect of environmental dynamism. Int. J. Prod. Econ. 2020, 222, 107498

Yakar, İ. D., & Taşlıyan, M. (2025). The rise of artificial intelligence and digital entrepreneurship in the Blue Ocean strategy: An evaluation of organizational outcomes. Nevşehir Hacı Bektaş Veli University Journal of Social

Sciences (Artificial Intelligence in Social Sciences : Theory, Application and Future Perspectives), 229-241.

Zeng, G., Ge, Y., Liu, L. (2013). From "innovation system" to "innovation ecosystem". Sci. Res., 1 pp. 4-11

Chapter 3

Few-Shot Learning: Conceptual Framework, Methodological Developments, and Security Dimensions

Sara Naghib Zadeh[1]

Hatice Nur Gök[2]

Abstract

The emergence of deep learning has largely been founded on the assumption of access to large-scale labeled datasets. However, in many real-world domains—ranging from medical imaging and the preservation of low-resource language families to space exploration and financial fraud detection—data are not always abundant or readily available. Annotation costs, privacy constraints, and the inherent rarity of certain phenomena constitute structural barriers to the practical deployment of machine learning systems. In this context, efforts to endow machines with the ability to "generalize from a few examples," a fundamental characteristic of human cognition, have given rise to the field of Few-Shot Learning (FSL).

This review article reconstructs the FSL literature around three principal paradigms: model fine-tuning, data augmentation, and transfer learning. Within the transfer learning framework, meta-learning mechanisms are examined in depth, and metric-based, optimization-based, and model-based approaches are comparatively analyzed. Furthermore, in light of the security threats that have recently emerged in this domain, feature-level adversarial attacks (FAMF) against metric-based models are comprehensively evaluated for the first time within this framework. The attack mechanisms, empirical findings, and defensive strategies are critically discussed. By identifying the structural challenges facing this field and outlining future research directions, this study argues that FSL should be regarded not merely as a technical subfield, but as a strategic turning point for the reliability and robustness of artificial intelligence systems.

1 Dr. Lecture, Halic University, Vocational School, Department of Computer Programming, ORCID: 0009-0005-6959-1165.

2 Halic University, Vocational School, Department of Big Data Analytics, ORCID:0009-0002-4248-6793

https://doi.org/10.58830/ozgur.pub1351.c5536

1. Introduction: The Problem of Data Scarcity in the Age of Data

Over the past decade, artificial intelligence research has been profoundly shaped by the deep learning revolution, which enabled the learning of hierarchical and multilayered representations. Landmark achievements in ImageNet-based image classification, breakthroughs in machine translation, and advances in speech modeling have all relied on massive quantities of labeled data and, consequently, substantial computational resources. However, the implicit assumption underlying this paradigm—that data are abundant and inexpensive—is increasingly being challenged (Parnami & Lee, 2021).

In many domains, data collection is an expensive and expertise-intensive process. For example, annotating histopathological images for pathological diagnosis requires hours of work by trained medical specialists. In the case of rare diseases, only a limited number of cases may exist worldwide. Similarly, the number of native speakers available for documenting indigenous languages is inherently restricted, and in space exploration, each collected sample may cost millions of dollars. In such scenarios, the recommendation to simply "collect more data" is often impractical and, in some cases, ethically unacceptable (Zhao et al., 2020).

At this point, human cognition serves as a major source of inspiration. A child can learn the concept of a "dog" after observing only a few examples, while an adult can recognize an unfamiliar individual in a crowded environment after seeing only a handful of photographs. This capability stems not merely from raw computational power, but from the transfer and integration of prior experiences into new situations. A person who knows how to ride a bicycle, for instance, can learn to ride a motorcycle more quickly because foundational skills such as balance and coordination are transferable (Parnami & Lee, 2021).

Few-Shot Learning (FSL) is precisely the field of research that seeks to systematically equip machines with this human-like capability for generalization. Formally, FSL refers to the ability of a model to correctly classify previously unseen classes in tasks where only a very limited number of training examples are available (typically $K = 1, 5,$ or 10) (Zhao et al., 2025). Learning from a single example is referred to as *One-Shot Learning*, while learning without any training examples and solely through auxiliary semantic information is known as Zero-Shot Learning (Pourpanah et al., 2022).

The contribution of this article is multifaceted. First, it reconstructs the existing FSL literature by organizing current approaches into three principal categories: model fine-tuning, data augmentation, and transfer learning, thereby providing a comprehensive conceptual map of the field. Second,

within the transfer learning framework, it comparatively analyzes meta-learning mechanisms from metric-based, optimization-based, and model-based perspectives. Third, for the first time in a review study of this scope, the security dimension of FSL is placed at the center of discussion through a detailed examination of Feature-level Adversarial Attacks on Metric-based Few-Shot Learning (FAMF). Finally, by identifying the structural challenges facing this field and outlining future research directions, the study proposes practical insights and recommendations for real-world applications (Zhao et al., 2025).

2.Theoretical Foundations of Few-Shot Learning

2.1. Formulation of the Few-Shot Classification Problem

In standard supervised learning, the objective is to train a model on a large dataset D in order to solve a specific task T. However, in the Few-Shot Learning (FSL) setting, the model is confronted with tasks for which the available training data are extremely limited. In the scientific literature, this problem is commonly formalized using the M -way K -shot framework:

M (number of classes): represents the number of novel categories that the model must distinguish.

K (number of examples): denotes the number of training samples available for each class during the learning phase.

Within this framework, the dataset is divided into two essential components:

Support Set: consists of K examples for each of the M classes, which the model uses for temporary adaptation and learning.

Query Set: contains unseen samples from the same classes, which the model must classify based on the knowledge acquired from the support set.

The fundamental distinction in FSL is that the model is trained on a set of "base" or "training" classes, yet during inference it is expected to recognize entirely new and previously unseen classes using only the limited information provided in the support set (Parnami & Lee, 2021).

2.2. Meta-Learning: The Mechanics of "Learning to Learn"

If classical deep learning can be described as the process of learning a specific skill, meta-learning may be understood as the process of learning how to learn. This concept, often regarded as the core driving mechanism of Few-Shot Learning, is inspired by the adaptive nature of human cognition.

2.2.1. Operational Principles of Meta-Learning

The primary objective of meta-learning is to expose the model to a wide variety of tasks rather than a single fixed problem. By solving thousands of small-scale tasks during the meta-training phase, the model learns which representations are generally transferable and how it can rapidly adapt to a new task with minimal parameter updates.

From a mathematical perspective, the goal is to identify optimal parameters θ that perform well not on a single task, but across a distribution of tasks $p(T)$:

$$\theta^* = arg\,min_\theta\, E_{\{T \sim p(T)\}}\left[L\left(D_{test}; f\left(D_{train}; \theta\right)\right)\right] \qquad (1)$$

In this formulation 1, L denotes the loss function evaluated on the test data D_{test}, while the model is allowed to adapt using only a very limited amount of training data D_{train} (Parnami & Lee, 2021).

Unlike conventional transfer learning, which primarily transfers knowledge from one domain to another, meta-learning aims to learn an adaptation strategy itself. This enables the model to achieve significantly higher robustness and adaptability when confronted with new tasks, even under conditions of limited labeled data or large-scale unlabeled environments, such as anomalous traffic patterns in distributed network systems.

2.2.2. Meta-Learning Taxonomy

Meta-learning approaches in Few-Shot Learning can be systematically categorized into three principal families based on the underlying mechanism of knowledge adaptation: optimization-based, model-based, and metric-based methods. Each family addresses the challenge of rapid adaptation from a distinct perspective, and their comparative characteristics are essential for understanding the methodological landscape of FSL.

Optimization-based meta-learning, exemplified by Model-Agnostic Meta-Learning (MAML) and its variants, seeks to learn an initialization of model parameters that can be fine-tuned with minimal gradient steps on new tasks. The core assumption is that a good initial parameter configuration exists in the loss landscape, from which rapid convergence to task-specific optima is possible (Finn et al., 2017).

Model-based meta-learning employs specialized architectures, such as memory-augmented neural networks or recurrent meta-learners, that explicitly store and retrieve task-relevant information. These models typically maintain an external memory module or use attention mechanisms to dynamically

adjust their internal state based on the support set, thereby enabling rapid task adaptation without extensive gradient-based optimization (Santoro et al., 2016; Munkhdalai & Yu, 2017).

Metric-based meta-learning, which will be examined in detail in Section 3.3.1, learns an embedding space where similarity between samples can be directly measured. Rather than adapting model parameters, these methods learn a fixed feature extractor and a distance metric, enabling classification by nearest-neighbor principles in the latent space (Snell et al., 2017; Sung et al., 2018).

The selection among these paradigms depends on the trade-off between computational cost, adaptation speed, and task complexity. While optimization-based methods offer greater flexibility, they incur higher computational overhead. Model-based approaches provide fast adaptation but require complex architectural designs. Metric-based methods excel in simplicity and inference speed but may lack discriminative power under extreme data scarcity (Parnami & Lee, 2021).

2.2.3. Meta-Learning in the Context of Few-Shot Learning

The integration of meta-learning into Few-Shot Learning represents a fundamental shift from conventional supervised learning paradigms. Whereas standard deep learning assumes access to abundant labeled data for each target class, meta-learning in FSL operates at the level of task distributions, enabling generalization across classes rather than within classes.

In the FSL setting, meta-learning is operationalized through episodic training. During each episode, a task is sampled consisting of a support set and a query set from a subset of classes. The model learns to minimize the classification error on the query set after being conditioned on the support set. Through thousands of such episodes, the model acquires meta-knowledge about how to extract discriminative features and construct decision boundaries from minimal examples.

This task-level learning distinguishes meta-learning from conventional transfer learning. In transfer learning, knowledge is transferred from source to target domains through shared representations or fine-tuned parameters. In meta-learning, the transfer occurs at the level of learning algorithms themselves—the model learns how to adapt, not merely what to transfer. This distinction is particularly critical in scenarios where target tasks are highly heterogeneous or where labeled data are too scarce to support effective fine-tuning (Parnami & Lee, 2021).

Furthermore, the meta-learning framework naturally accommodates the M-way K-shot formulation of FSL. The episodic training protocol ensures that the model is explicitly optimized for the few-shot scenario, rather than being implicitly expected to generalize from limited data after standard supervised pretraining. This alignment between training and evaluation conditions constitutes one of the primary reasons for the superior empirical performance of meta-learning-based FSL methods compared to simple fine-tuning approaches.

3. Few-Shot Learning Approaches: Three Core Paradigms

Following the establishment of the theoretical framework of Few-Shot Learning (FSL), researchers have sought to develop methods capable of addressing the fundamental challenge of data scarcity in real-world environments. Unlike conventional deep learning approaches, which rely on large-scale annotated datasets, FSL requires the design of mechanisms that enable models to generalize from only a handful of training examples. In this context, the literature has gradually been organized around three main paradigms: (1) fine-tuning-based approaches, (2) data augmentation-based approaches, and (3) transfer learning and meta-learning-based approaches. Each of these paradigms addresses the data scarcity problem from a different perspective and aims to balance computational complexity, accuracy, generalization ability, and training cost (Parnami & Lee, 2021). A comparative overview of these paradigms is presented in Table 3-1.

Table 3-1. Comparison of the Main Paradigms in Few-Shot Learning (FSL)

Paradigm	Core Idea	Advantages	Limitations	Representative Studies
Fine-Tuning-Based Approaches	Transfer knowledge from pretrained models to low-data tasks by updating model parameters	Simple implementation, leverages prior knowledge, reduces need for training from scratch	Sensitive to domain shift, prone to overfitting in low-data regimes	ULMFiT (Howard & Ruder, 2018), Nakamura et al. (2019)
Data Augmentation-Based Approaches	Enrich datasets using unlabeled data, synthetic samples, or feature augmentation	Improves data diversity, enhances generalization, mitigates data scarcity	Difficulty in modeling true data distribution, limited generalization to unseen classes	Wang et al. (2016), Mehrotra et al. (2017), Chen et al. (2019)
Transfer / Meta-Learning-Based Approaches	Learn fast adaptation mechanisms across a distribution of tasks	High generalization, fast adaptation, suitable for dynamic environments	High computational complexity, difficult optimization	Finn et al. (2017), Santoro et al. (2016), Garcia & Bruna (2018)

3.1. Fine-Tuning-Based Approaches: Knowledge Transfer from Pretrained Models

Fine-tuning-based approaches represent the most traditional and direct strategy in Few-Shot Learning. The central idea is that a deep neural network pretrained on a large-scale and general-purpose dataset can transfer the representations learned in its layers to a new task with limited data. In this paradigm, the initial model is typically trained on datasets such as ImageNet, and then adapted using only a small number of samples from the target task. The underlying assumption is that early-layer representations capture general-purpose features that can be effectively reused across different domains (Parnami & Lee, 2021).

One of the most influential approaches in transfer learning for low-data settings is ULMFiT, introduced by Jeremy Howard and Sebastian Ruder (2018). Unlike earlier methods that only retrained the final classification layer, ULMFiT proposed a three-stage framework consisting of general language model pretraining, domain-specific fine-tuning, and task-specific classification training. The model also introduced layer-wise learning rates, where earlier layers are updated more conservatively and later layers more aggressively, enabling the preservation of general knowledge while adapting to new tasks. In addition, the use of slanted triangular learning rates, which initially increase and then gradually decrease during training, helps accelerate convergence and reduce overfitting, making the approach particularly effective in Few-Shot Learning scenarios with limited training data (Howard & Ruder, 2018).

Nakamura et al. (2019) further proposed a similar strategy, employing lower learning rates for low-shot classes alongside adaptive gradient-based optimization methods. The main objective was to prevent drastic parameter updates when training with scarce data. However, despite their simplicity and ease of implementation, fine-tuning-based approaches suffer from a key limitation: when there is a significant domain gap between source and target datasets, the risk of overfitting increases substantially, leading to poor generalization performance. A comparative summary of representative fine-tuning methods is presented in Table 3-2. These limitations have motivated the development of more advanced approaches based on transfer learning and meta-learning (Nakamura et al., 2019).

Table 3-2. Comparison of Representative Fine-Tuning Methods

Model / Method	Core Idea	Key Innovation	Limitation	Reference
ULMFiT	Fine-tuning a pretrained language model for low-resource tasks	Layer-wise learning rates and slanted triangular learning rates	Sensitive to domain shift	Howard & Ruder (2018)
Adaptive Fine-Tuning	Using lower learning rates for low-shot classes	Adaptive gradient-based optimization	Overfitting under domain mismatch	Nakamura et al. (2019)

3.2. Data Augmentation Approaches: Enriching the Learning Space

One of the most critical challenges in Few-Shot Learning (FSL) is the statistical scarcity arising from the limited number of training samples. When a model is exposed to only a small number of examples, it cannot adequately learn the true diversity of the underlying data distribution, which consequently reduces its generalization capability. Data augmentation approaches have been developed to address this limitation by enriching small datasets, either through generating new data or leveraging auxiliary information to expand the learning space (Parnami & Lee, 2021). A general categorization of these methods, along with their key characteristics, is presented in Table 3-3.

Table 3-3. Categorization of data augmentation methods in FSL

Subcategory	Core Idea	Representative Models / Studies	Advantages	Limitations
Unlabeled data utilization	Using unlabeled data to learn general representations	Wang et al. (2016), Boney et al. (2018)	Reduces dependence on labeled data	Sensitive to representation quality
Transductive learning	Incorporating test data structure into the learning process	TPN — Liu et al. (2019)	Improves generalization	Increased computational cost
Data synthesis	Generating synthetic samples for low-shot classes	Mehrotra et al. (2017), Hariharan & Girshick (2017), Wang et al. (2018)	Increases training diversity	Difficulty in modeling true data distribution
Feature augmentation	Enriching feature space instead of sample space	Dixit et al. (2017), Liu et al. (2018), Schwartz et al. (2018), Chen et al. (2019)	Produces richer representations	Limited interpretability of features

3.2.1. Use of Unlabeled Data

In many real-world applications, unlabeled data are significantly more abundant than labeled data. Therefore, one of the earliest strategies was to exploit unlabeled data to enhance model learning. Wang et al. (2016) proposed a transfer-based approach using convolutional neural networks, introducing a self-supervised pre-training phase. In this stage, the model learns a general and rich representation of the data space without being constrained to specific classes. This representation is subsequently used in few-shot tasks (Wang et al., 2016).

Similarly, Boney et al. (2018) introduced a semi-supervised variant of MAML, in which unlabeled data are used to refine the embedding function, while labeled data are used to train the classifier. This combination enables the model to learn more stable representations even under extreme data scarcity (Boney et al., 2018).

3.2.2. Transductive Learning

Transductive learning can be considered a subset of semi-supervised learning in which the structure of test data is incorporated into the learning process. Unlike inductive learning, which assumes test data are completely unseen, transductive learning jointly analyzes relationships between training and test samples.

In this context, Liu et al. (2019) proposed the Transductive Propagation Network (TPN). This model consists of four main stages: feature extraction, graph construction, label propagation, and loss computation. The key idea of TPN is that the geometric structure of unlabeled test data can help the model construct more accurate decision boundaries (Liu et al., 2019).

3.2.3. Data Synthesis and Generative Networks

Another important approach to data augmentation is the generation of synthetic samples for low-shot classes. Generative Adversarial Networks (GANs), due to their strong capability in modeling data distributions, have become a central tool in this domain.

Mehrotra et al. (2017) proposed a GAN-based architecture for One-Shot Learning. In this framework, the generator aims to produce synthetic samples that are close to the true data distribution (Mehrotra et al., 2017). Hariharan and Girshick (2017) introduced a two-stage approach consisting of representation learning and multi-shot classification, where synthetic data are used to improve classification performance (Hariharan & Girshick, 2017).

Furthermore, Wang et al. (2018) developed a model that integrates meta-learning with data generation to produce virtual samples for novel classes .

Despite their success, data synthesis methods still face significant challenges. Many generative models struggle to accurately capture complex data distributions, resulting in synthetic samples that may lack realistic structure. Moreover, their generalization to entirely novel classes remains difficult, and generated features often have limited interpretability (Parnami & Lee, 2021).

3.2.4. Feature Augmentation

Instead of directly generating samples, some researchers have focused on enriching the feature space. In this approach, the goal is to enable the model to learn intrinsic variations within data at the representation level.

Shu introduced the AGA model, which manipulates images based on object-level features (Shu et al., 2018). Schwartz et al. (2018) proposed the Delta Encoder, demonstrating that new features for unseen classes can be synthesized even from a few examples (Schwartz et al., 2022).

In addition, Chen et al. (2019) introduced TriNet, which establishes a bidirectional mapping between the semantic label space and the image feature space. This design allows each class to obtain a richer and more robust representation in the feature space (Chen et al., 2019).

3.3. Transfer Learning and Meta-Learning Approaches

3.3.1. Metric-Based Methods: Learning the Concept of Similarity

In many Few-Shot Learning (FSL) scenarios, the number of available samples is so limited that training a deep classifier directly becomes practically infeasible. Consequently, researchers have shifted from directly learning class labels toward learning the concept of *similarity*. Metric learning is based on the assumption that samples belonging to the same class should be close to each other in the feature space, while samples from different classes should be far apart (Parnami & Lee, 2021).

From a mathematical perspective, a metric is a function that measures the distance between two samples. In deep learning settings, commonly used metrics include Euclidean distance, Mahalanobis distance, and cosine similarity. A typical metric learning framework consists of two components: an embedding module that maps input data into a vector space, and a metric module that computes similarity or distance between embeddings (Kotovenko et al., 2023).The most important metric-based models and their characteristics are compared in Table 3-4.

Table 3-4. Comparison of metric-based learning models

Model	Core Idea	Distinct Feature	Advantages	Reference
Siamese Networks	Learning similarity between pairs of samples	Shared weights between twin networks	Suitable for one-shot learning	Koch et al. (2015)
Matching Networks	Using attention for sample comparison	LSTM + attention mechanism	High accuracy in low-shot settings	Vinyals et al. (2016)
Prototypical Networks	Defining a prototype for each class	Class-wise feature averaging	Simplicity and fast inference	Snell et al. (2017)
Relation Networks	Learning comparison function via CNN	Eliminates need for predefined metric	High flexibility	Sung et al. (2018)
CovaMNet	Using feature covariance information	Second-order statistical representation	Richer feature representation	Li et al. (2019)

Koch et al. (2015) introduced Siamese Neural Networks for one-shot recognition. These networks consist of two identical branches with shared weights and aim to minimize the distance between similar samples while maximizing the distance between dissimilar ones (Koch et al., 2015). Subsequently, Vinyals et al. (2016) proposed Matching Networks, which learn similarity using LSTM and attention mechanisms (Vinyals et al., 2016).

Snell et al. (2017) introduced Prototypical Networks, where each class is represented by a prototype (i.e., the mean of its feature embeddings), and classification is performed based on the nearest prototype in feature space (Snell et al., 2017). Later, Sung et al. (2018) proposed Relation Networks, which learn the similarity function directly using a neural network instead of relying on predefined distance metrics (Sung et al., 2018).

Additionally, Li et al. (2019) introduced CovaMNet, which leverages covariance matrices to capture second-order statistical information, resulting in richer representations compared to mean-based approaches (Li et al., 2019).

The main advantages of metric-based methods include computational simplicity, fast inference, and rapid adaptation to new tasks. However, under extremely limited data conditions, simple distance metrics may lack sufficient discriminative power, which can reduce performance in dynamic environments (Parnami & Lee, 2021).

3.3.2. Meta-Learning Methods: Learning to Learn

Meta-learning, or "learning to learn," is one of the most fundamental frameworks in Few-Shot Learning. Unlike conventional machine learning, where a model is trained for a single task, meta-learning aims to enable models to rapidly adapt to new tasks with minimal data (Parnami & Lee, 2021).

In this framework, data are divided into two levels: meta-training and meta-testing. Each task consists of a support set and a query set. The model is trained over a distribution of tasks so that it can generalize to unseen tasks with only a few examples. A comparison of key meta-learning models is provided in Table 3-5.

Table 3-5. Comparison of meta-learning methods

Model	Architecture / Core Idea	Key Advantage	Limitation	Reference
MANN	External memory with neural Turing machine	Fast learning from limited data	High architectural complexity	Santoro et al. (2016)
Meta Networks	Combination of meta-learner and base learner	Rapid weight generation	High training cost	Munkhdalai & Yu (2017)
MAML	Learning adaptable initial parameters	Model-agnostic approach	Requires second-order gradients	Finn et al. (2017)
TAML	Regularization to reduce overfitting	Better generalization	Difficult hyperparameter tuning	Jamal et al. (2019)

One of the earliest influential models in this area is Memory-Augmented Neural Networks (MANN), introduced by Santoro et al. (2016). This model integrates external memory with a neural Turing machine architecture, enabling both short-term and long-term information storage (Santoro et al., 2016).

Munkhdalai and Yu (2017) proposed Meta Networks, consisting of a base learner and a meta-learner. The meta-learner generates fast weights that allow rapid adaptation to new tasks (Munkhdalai & Yu, 2017).

Finn et al. (2017) introduced Model-Agnostic Meta-Learning (MAML), one of the most influential meta-learning approaches. The core idea is to learn an initialization of model parameters that can be quickly adapted to new tasks using only a few gradient steps:

$$\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{T_i}\left(f_\theta\right) \qquad (2)$$

The main strength of MAML lies in its model-agnostic nature, meaning it can be applied to any gradient-based model. However, its reliance on second-order derivatives increases computational cost. To address this, variants such as FOMAML and Reptile were proposed (Finn et al., 2017).

Jamal et al. (2019) further proposed TAML, which introduces a regularization term to reduce overfitting on training tasks and improve generalization performance (Jamal et al., 2019).

3.3.3. Graph Neural Network-Based Methods: Modeling Structural Relationships

Graph Neural Networks (GNNs) represent another important direction in Few-Shot Learning, designed to model structural relationships among samples. In this approach, data are represented as graphs, where each sample is a node and relationships between samples are represented as edges (Parnami & Lee, 2021).

Garcia and Bruna (2018) proposed one of the first GNN-based models for few-shot image classification. In this architecture, the model jointly learns node and edge representations and explicitly models relationships among samples (Garcia & Bruna, 2018).

Later, Kim et al. (2019) developed EGNN, which focuses on edge classification. In this model, each edge is represented by a two-dimensional feature vector indicating whether two nodes belong to the same class or different classes (Kim et al., 2019). A comparison of these models is provided in Table 3-6.

Table 3-6. Comparison of GNN-based methods in FSL

Model	Core Idea	Key Feature	Advantages	Limitation	Reference
GNN	Modeling samples as graphs	Joint node-edge learning	High interpretability	Complexity grows with graph size	Garcia & Bruna (2018)
EGNN	Focus on edge classification	Probabilistic edge representation	Captures complex relations	High computational cost	Kim et al. (2019)

The main advantage of GNN-based methods is their strong interpretability and ability to model complex dependencies among samples. However, as the

number of samples increases, the number of edges grows exponentially, leading to significant computational overhead (Parnami & Lee, 2021).

Finally, a comparative summary of the main FSL families in terms of computational complexity, generalization capability, and robustness to overfitting is presented in Table 3-7.

Table 3-7. Comparative summary of major FSL families

Method Family	Computational Complexity	Inference Speed	Overfitting Resistance	Generalization	Interpretability
Model Fine-Tuning	Medium	High	Medium	Domain-dependent	Medium
Data Augmentation	Medium–High	Medium	High	Relatively good	Low
Metric Learning	Low	Very high	Medium	Good	High
Meta-Learning	High	High	High	Very strong	Medium
Graph Neural Networks	High	Medium	High	Good	High

3.4. Integrated Conceptual Framework: Synthesis of FSL Paradigms

The three paradigms examined in the preceding sections—fine-tuning, data augmentation, and transfer/meta-learning—are not mutually exclusive methodological alternatives, but rather complementary strategies that can be integrated to address the multifaceted challenge of data scarcity. This section presents an integrated conceptual framework that synthesizes these paradigms and clarifies their interrelationships within the broader FSL ecosystem.

At the foundational level, fine-tuning-based approaches leverage pretrained representations as a starting point for adaptation. This paradigm is most effective when the source and target domains share substantial visual or semantic similarity, and when computational resources are limited. However, as the domain gap widens or data scarcity becomes more severe, fine-tuning alone proves insufficient, necessitating the incorporation of additional mechanisms.

Data augmentation operates at the sample level, enriching the training space through synthetic or transformed examples. When combined with fine-tuning, augmentation mitigates overfitting and enhances generalization. More critically, when integrated with meta-learning, augmentation can improve the diversity of episodic tasks, thereby strengthening the model's ability

to generalize across task distributions. For instance, feature augmentation strategies can be embedded within meta-learning episodes to provide richer support sets without requiring additional labeled data (Chen et al., 2019).

Transfer and meta-learning represent the highest level of abstraction in this framework. Meta-learning learns the adaptation mechanism itself, while transfer learning provides the representational substrate upon which this mechanism operates. The synergy between these approaches is evident in methods such as MAML, where pretrained initializations can accelerate meta-training, and in metric-based approaches, where transfer-learned embeddings serve as the feature space for similarity computation.

The selection of an appropriate strategy—or combination thereof—depends on several contextual factors: (i) the degree of domain shift between source and target data; (ii) the availability of unlabeled or auxiliary data; (iii) computational constraints; and (iv) the required adaptation speed. In practice, hybrid architectures that combine meta-learning with data augmentation and transfer learning have demonstrated superior performance over single-paradigm approaches, suggesting that the future of FSL lies in the principled integration of these complementary strategies rather than in their isolated application (Parnami & Lee, 2021; Song et al., 2023).

Figure 3-1 illustrates the proposed integrated framework, depicting the hierarchical relationships among the three paradigms and their potential intersections. The framework emphasizes that effective FSL systems should be designed with flexibility to incorporate multiple paradigms, adapting their composition to the specific requirements of the target application domain.

4. Experimental Evaluation and Performance Outlook in Few-Shot Learning

Experimental evaluation plays a central role in assessing the generalization capability of Few-Shot Learning (FSL) models. Unlike traditional supervised learning, where models are evaluated on large-scale data, FSL focuses on rapid adaptation to novel classes with extremely limited labeled samples. For this reason, standardized benchmark datasets such as miniImageNet and Omniglot have become widely used for fair comparison of models (Zhao et al., 2024).

Omniglot consists of handwritten characters from multiple alphabets and is considered a relatively simple benchmark due to its limited visual variability. In contrast, miniImageNet contains real-world images with significant variations in background, illumination, viewpoint, and object structure, making it a much more challenging benchmark for evaluating generalization performance (Parnami & Lee, 2021).

Most studies report results under standard evaluation settings such as 5-way 1-shot and 5-way 5-shot classification. In the 1-shot setting, the model observes only one example per class, whereas in the 5-shot setting it observes five examples per class, allowing a more robust estimation of class structure (Xian et al., 2020).

4.1. Empirical Results on miniImageNet

Table 4-1 summarizes the performance of representative FSL models on miniImageNet under 1-shot and 5-shot settings.

Table 4-1. Performance comparison of FSL models on miniImageNet (Parnami & Lee, 2021).

Model	Approach Type	1-shot Accuracy (%)	5-shot Accuracy (%)
Matching Networks	Metric learning	43.56	55.31
Prototypical Networks	Metric learning	49.42	68.20
MAML	Optimization-based meta-learning	48.70	63.11
Relation Networks	Metric learning	50.44	65.32
SimpleShot	Transfer learning	64.29	80.64

Several important observations can be drawn from these results. First, increasing the number of support samples from 1-shot to 5-shot consistently improves performance across all models. This indicates that even a small increase in labeled examples significantly enhances class representation quality. For instance, Prototypical Networks improve from 49.42% to 68.20% accuracy (Snell et al., 2017).

Studies in Few-Shot Learning consistently show a clear performance gap between simpler benchmarks such as Omniglot and more complex datasets like miniImageNet, where accuracy drops significantly due to higher visual diversity, emphasizing the importance of robust and transferable feature representations. In addition, results across methods indicate a steady improvement over time, with early approaches like Matching Networks achieving around 43% accuracy in the 1-shot setting, while later methods such as SimpleShot surpass 64%, reflecting advances in representation learning strategies. Notably, SimpleShot demonstrates that even without complex meta-learning mechanisms, strong

performance can be achieved through high-quality feature embeddings and proper normalization, sometimes outperforming more sophisticated models (Parnami & Lee, 2021).

4.2. Structural Perspective on Performance Trends

Empirical evidence shows that FSL performance is not solely determined by model complexity. Instead, it depends on multiple interacting factors, including feature representation quality, similarity metric design, task construction strategy, and domain shift between training and testing distributions.

Metric-based methods generally perform well due to their simplicity and fast inference. However, they may struggle in highly complex or noisy environments where simple distance measures are insufficient. In contrast, meta-learning approaches provide stronger adaptation capabilities but require higher computational cost and carefully designed task distributions for training (Finn et al., 2017).

Moreover, the success of models such as SimpleShot indicates that the quality of the embedding space can be more critical than architectural complexity. This observation has led to an increasing research focus on representation learning rather than purely on adaptation mechanisms (Payandeh et al., 2023).

Overall, experimental studies suggest that no single FSL approach is universally optimal. The best-performing method depends on dataset characteristics, task complexity, computational constraints, and the degree of domain shift between training and testing environments.

5. Security in Few-Shot Learning: Adversarial Attacks and Model Vulnerability

5.1. Feature-Level Adversarial Attacks and Model Vulnerability

Deep learning models are highly vulnerable to adversarial attacks, as it has been shown that even very small perturbations in the input can lead to significant decision-making errors (Goodfellow et al., 2015). More advanced methods such as PGD and Carlini–Wagner attacks further revealed and intensified this vulnerability in neural networks (Madry et al., 2018). This issue becomes even more critical in Few-Shot Learning (FSL), where models are trained with extremely limited samples, making their decision boundaries more fragile and highly sensitive to small input variations. For this reason, methods such as ADML, AQ, and DFSL have been proposed to improve robustness, although the security challenge has not yet been fully resolved (Kim et al., 2026).

In classical adversarial attacks, a small perturbation is added to the input:

$$x^{adv} = x + \delta s.t. \square \delta \square \leq \grave{o}$$

This approach is effective in image-based classification models, but it has limitations in metric-based Few-Shot models, since these models make decisions based on distances in the feature space rather than the pixel space. Therefore, effective attacks require manipulation at the feature level rather than the input level (Kim et al., 2026 ; Xu et al., 2025).

In this context, Kim et al. (2026) introduced FAMF (Feature-level Adversarial Attack on Metric-based Few-Shot Learning), the first attack specifically designed for metric-based FSL models. In this method, the perturbation is applied directly in the feature space:

$$f(x)^{adv} = f\left(x\right) + \arg\max_{\delta} L\left(\theta, f\left(x\right) + \delta, y\right) s.t. \square \delta \square \leq \grave{o}$$

The goal of this attack is to increase classification error by reducing intra-class similarity and increasing inter-class similarity in the feature space (Kim et al., 2026).

Experimental results on Omniglot and miniImageNet datasets show that FAMF is significantly more effective than traditional attacks such as PGD, achieving nearly 100% attack success rate in some models like FEAT. In the 1-shot setting, attacks are stronger than in the 5-shot setting because fewer support samples make decision boundaries more fragile. Moreover, FAMF remains effective even in the presence of defense mechanisms such as AQ, and it shows higher robustness against defensive strategies compared to image-level attacks. From a computational perspective, FAMF is also faster than PGD, since it operates directly in the feature space and does not require full backpropagation through the entire network (Kim et al., 2026 ; Xu et al., 2025).

These findings indicate that the main vulnerability of metric-based Few-Shot models lies in the feature representation space, rather than the input space. Therefore, the security of these models must be redefined beyond input-level defenses and extended toward feature-level defense mechanisms, which ensure the robustness of the latent representation space against adversarial perturbations. Additionally, FAMF can be used as a diagnostic tool to identify hidden weaknesses in models and to design more robust architectures (Zheng et al., 2025).

However, this type of attack is mainly developed under a white-box assumption, where the attacker has full access to the model architecture and embedding function. Therefore, developing black-box versions of such attacks

and designing robust defenses against feature-level adversarial strategies remain among the most important future research directions in the security of Few-Shot Learning (Cao et al., 2020 ; Kim et al., 2026).

5.2. Poisoning Attacks in Few-Shot Learning

Beyond adversarial perturbations at inference time, Few-Shot Learning models are vulnerable to poisoning attacks that compromise the integrity of the training or support data. In the FSL context, poisoning can occur at two distinct stages: during meta-training, where the attacker contaminates the base dataset used to learn the meta-parameters, or during inference, where malicious examples are injected into the support set.

Meta-training poisoning is particularly insidious because the attack is embedded in the model's learned prior. By strategically inserting mislabeled or crafted examples into the training tasks, an attacker can bias the meta-learner toward representations that fail under specific trigger conditions. For instance, a backdoored meta-model might perform normally on standard few-shot tasks but produce systematically incorrect predictions when a predefined visual pattern appears in the query image (Gu et al., 2019).

Support set poisoning targets the adaptation phase of FSL. Since metric-based and meta-learning models rely heavily on the support set for task-specific decision boundaries, even a small number of poisoned support examples can significantly degrade classification accuracy. Unlike standard supervised learning, where poisoning effects may be diluted across a large training set, FSL's extreme data scarcity amplifies the impact of each corrupted sample. Recent studies have demonstrated that injecting as few as one poisoned example per class into the support set can reduce accuracy by over 30% in prototypical networks (Shafahi et al., 2018).

The defense against poisoning attacks in FSL requires robust data validation mechanisms at both the meta-training and inference stages. Techniques such as spectral signature detection, activation clustering, and outlier removal have shown promise in identifying poisoned samples, though their adaptation to the few-shot setting remains an active area of research.

5.3. Backdoor Attacks

Backdoor attacks represent a specialized form of poisoning in which the attacker embeds a hidden trigger pattern into the model during training, causing the model to behave normally on clean inputs but produce attacker-chosen outputs when the trigger is present. In Few-Shot Learning, backdoor

attacks are especially dangerous due to the limited number of training examples, which makes it easier to introduce trigger patterns without detection.

In metric-based FSL, backdoor attacks can be implemented by manipulating the embedding space such that trigger-embedded samples from any class map to a specific region associated with a target class. During inference, any query image containing the trigger will be misclassified into the attacker's chosen category, regardless of its true class. This attack is particularly effective against prototypical networks, where the class prototype can be shifted toward the trigger-embedded region through a small number of poisoned support examples (Liu et al., 2020).

The stealthiness of backdoor attacks in FSL stems from the fact that clean task performance remains largely unaffected, making detection through standard accuracy evaluation difficult. Furthermore, the episodic training protocol of meta-learning provides additional cover for the attacker, as the trigger can be distributed across multiple tasks without appearing suspicious in any single episode.

Defensive strategies against backdoor attacks in FSL include neural cleanse techniques, which reverse-engineer potential triggers by analyzing model behavior across classes, and fine-pruning methods that remove dormant neurons associated with backdoor functionality. However, these defenses were originally designed for standard supervised learning and require significant adaptation for the meta-learning setting, where model parameters are optimized for rapid task adaptation rather than fixed classification boundaries.

5.4. Membership Inference Attacks

Membership inference attacks aim to determine whether a specific data sample was included in the training set of a machine learning model. While traditionally studied in the context of large-scale supervised learning, membership inference poses distinct challenges and risks in Few-Shot Learning due to the intimate relationship between support set composition and model predictions.

In FSL, membership inference can be targeted at two levels: the meta-training set and the task-level support set. At the meta-training level, an attacker with query access to the trained meta-model can infer whether a particular class or sample was included in the meta-training distribution. This is particularly concerning in applications involving sensitive data, such as medical diagnosis or biometric identification, where the mere presence of an individual's data in the training set may constitute a privacy violation (Shokri et al., 2017).

At the support set level, membership inference becomes even more direct. Since the support set is explicitly used to condition the model's predictions during inference, an attacker can exploit the model's confidence patterns to infer which samples were present in the support set. For example, in prototypical networks, the distance between a query sample and its corresponding class prototype tends to be smaller when the query was part of the support set, providing a discriminative signal for membership inference.

The vulnerability of FSL models to membership inference is exacerbated by their reliance on distance metrics in the embedding space. These metrics often leak information about the support set composition, particularly when the number of support examples is extremely small. Defense mechanisms such as differential privacy, which adds calibrated noise to model outputs or gradients, have been proposed to mitigate membership inference risks. However, applying differential privacy in FSL is challenging because the noise may degrade the already limited information available for task adaptation, creating a tension between privacy preservation and few-shot performance.

5.5. Model Extraction Attacks

Model extraction attacks involve an adversary with only black-box query access to a target model, attempting to construct a functionally equivalent copy of that model. In Few-Shot Learning, this threat is particularly acute because FSL models are often deployed as lightweight, specialized services where query access is provided to users for rapid task adaptation.

The extraction process in FSL differs from standard model extraction due to the episodic nature of inference. An attacker can query the target model with carefully constructed support sets and query samples, observing the predicted labels or confidence scores. By systematically exploring the input space across multiple episodes, the attacker can train a surrogate model that approximates the target model's embedding function and decision boundaries.

Metric-based FSL models are especially susceptible to extraction attacks because their decision logic is relatively simple: compute embeddings, measure distances, and select the nearest class. This simplicity enables accurate extraction with fewer queries compared to complex deep classifiers. Furthermore, because FSL models are designed for rapid adaptation, they often lack the defensive depth—such as input preprocessing pipelines or ensemble structures—that might otherwise impede extraction (Tramer et al., 2016).

The consequences of successful model extraction in FSL extend beyond intellectual property theft. An extracted surrogate model can be used to mount more effective white-box attacks, including the feature-level adversarial

attacks and backdoor injections described in previous sections. Moreover, if the original model was trained on proprietary or sensitive data, the extracted model may retain traces of that data, creating secondary privacy risks.

Defenses against model extraction in FSL include rate limiting on query access, output perturbation to obscure confidence scores, and watermarking techniques that embed identifiable signatures in model predictions. However, these defenses must be carefully calibrated to avoid degrading the core few-shot adaptation capability that makes FSL valuable in the first place.

5.6. Comprehensive Defense Strategies

The diverse attack surface of Few-Shot Learning—spanning feature-level adversarial perturbations, data poisoning, backdoor triggers, membership inference, and model extraction—necessitates a multi-layered defense strategy that addresses vulnerabilities at each stage of the FSL pipeline. This section synthesizes the defensive approaches discussed throughout Section 5 and proposes an integrated security framework for robust FSL deployment.

At the input level, adversarial training remains the most effective defense against feature-level attacks such as FAMF. By augmenting the meta-training process with adversarially perturbed support and query sets, the model learns to construct more robust decision boundaries in the embedding space. However, standard adversarial training significantly increases computational cost and may reduce clean-task accuracy, necessitating the development of efficient adversarial training variants tailored to the episodic learning paradigm (Madry et al., 2018).

At the data level, robust aggregation and outlier detection mechanisms can mitigate poisoning and backdoor attacks. For meta-training, spectral analysis of gradient updates across tasks can identify and exclude poisoned episodes. For inference-time support sets, consistency checks based on geometric properties of the embedding space—such as unexpected prototype shifts or anomalous inter-sample distances—can flag potentially corrupted support examples before they influence predictions.

At the model level, architectural modifications can enhance intrinsic robustness. Ensemble approaches that aggregate predictions from multiple meta-learners with diverse initializations reduce the impact of any single compromised component. Additionally, regularization techniques that enforce smoothness in the embedding space—such as Lipschitz continuity constraints—limit the adversary's ability to induce large changes in model output through small perturbations.

At the system level, access control and monitoring mechanisms are essential for preventing model extraction and membership inference. Query logging, anomaly detection in query patterns, and differential privacy guarantees can collectively raise the cost of attacks while preserving legitimate user functionality. The trade-off between security and usability must be carefully managed, as excessive restrictions may undermine the rapid adaptation capability that defines FSL.

The relationship between attack vectors and corresponding defenses can be summarized as follows. Feature-level adversarial attacks such as FAMF are primarily countered through adversarial training at the input level, with embedding space regularization serving as a complementary secondary defense. Poisoning attacks targeting either the meta-training set or the support set require spectral signature detection as the primary defense, supplemented by outlier removal techniques. Backdoor attacks embedded during meta-training are addressed through neural cleanse methods, with fine-pruning of dormant neurons as an additional safeguard. Membership inference attacks at the inference stage are mitigated through differential privacy mechanisms, supported by output perturbation to obscure confidence patterns. Finally, model extraction attacks at the system level are countered through rate limiting and query logging, with watermarking techniques providing secondary protection for intellectual property.

In conclusion, the security of Few-Shot Learning systems cannot be ensured through any single defensive mechanism. Rather, a defense-in-depth approach that combines input sanitization, robust training, architectural hardening, and system-level monitoring is required to address the multifaceted threat landscape. As FSL continues to be deployed in safety-critical and privacy-sensitive applications, the integration of security considerations into the core design of FSL methodologies will become not merely advisable but essential.

6. Challenges and Future Directions in Few-Shot Learning

Despite significant advances in Few-Shot Learning (FSL), the field still faces several fundamental challenges that limit its full applicability in real-world scenarios. One of the most important limitations is the reliance on the conventional **M-way K-shot** setting, where models are trained under highly controlled conditions. In real-world problems, however, the number of classes and the number of samples per class are not predefined, which reduces model flexibility and adaptability. Moreover, meta-learning approaches often assume that training and testing tasks are independently and identically distributed according to a task distribution $p(T)$. This assumption leads to significant

performance degradation under cross-domain scenarios, such as transferring from natural images to text or audio data (Parnami & Lee, 2021).

Another key challenge is the difficulty of integrating knowledge from both seen and unseen classes within a unified framework. Many existing models are designed only to classify novel classes within a fixed support set, which limits their effectiveness in real-world environments where both old and new classes must be recognized simultaneously. In addition, data heterogeneity in domains such as audio, wireless signals, and text prevents the construction of standardized datasets, which are essential for stable meta-learning. Furthermore, the vulnerability of FSL models to adversarial attacks and feature-space manipulation raises serious concerns regarding robustness and trustworthiness (Kim et al., 2026).

To address these challenges, future research directions focus on several key areas. First, the integration of structured prior knowledge, such as knowledge graphs and ontologies, may reduce reliance on large-scale pretraining data. Second, the development of task-adaptive distance metrics, which can dynamically adjust instead of relying on fixed measures such as Euclidean distance, is considered a promising direction for improving model flexibility.

In addition, advanced meta-learning architectures, particularly hierarchical models, can enhance the ability to capture meta-knowledge and better handle task heterogeneity. The combination of different learning paradigms—including transfer learning, active learning, and reinforcement learning—may also lead to more powerful hybrid systems. Furthermore, multi-modal learning approaches that enable knowledge transfer across text, image, and audio modalities, as well as progress in zero-shot learning, represent important future research directions (Da Silva & Costa, 2019).

Finally, robustness and security remain critical concerns. The development of feature-level defense mechanisms, the use of adversarial training to improve robustness, and the adaptation of certified robustness concepts to FSL settings are essential strategies for ensuring reliability in safety-critical applications (Parnami & Lee, 2021; Kim et al., 2026).

7. Conclusion

7.1. Summary and Key Contributions

Few-Shot Learning (FSL) has emerged as one of the most important research directions in machine learning, particularly in scenarios where data scarcity is a fundamental limitation. This review systematically organized the FSL literature into three main paradigms: model fine-tuning, data augmentation,

and transfer learning. Within the transfer learning paradigm, meta-learning approaches—including metric-based, optimization-based, and model-based methods—were comparatively analyzed, highlighting their strengths and limitations (Parnami & Lee, 2021).

A key contribution of this study is the emphasis on security and robustness alongside predictive performance. Results from feature-level adversarial attacks such as FAMF (Feature-level Adversarial Attack) demonstrate that while metric-based models perform well at the input level, they remain vulnerable in the feature representation space. This indicates that FSL systems should not only focus on improving accuracy but must also explicitly consider robustness, stability, and trustworthiness in their design (Kim et al., 2026).

Moreover, several structural challenges remain unresolved, including the limitations of the M-way K-shot framework, dependence on fixed task distributions, difficulty in integrating seen and unseen classes, and poor generalization to non-visual domains. These challenges suggest that FSL is still an evolving field requiring fundamental innovations. Promising future directions include the use of prior knowledge, development of adaptive distance metrics, integration of multiple learning paradigms, and design of robust models resistant to adversarial attacks (Song et al., 2023).

In conclusion, Few-Shot Learning is not merely a technical machine learning approach but a strategic research area that will play a crucial role in the future of artificial intelligence. In a world where data scarcity is a universal constraint, successful advancement of FSL can enable broader deployment of AI systems in real-world, low-data, and complex environments, ultimately contributing to more accessible and equitable intelligent technologies (Song et al., 2023).

7.2. Foundation Models and the Evolution of Few-Shot Learning

The emergence of large-scale foundation models—such as GPT-4, CLIP, DINO, and Segment Anything Model (SAM)—has fundamentally altered the landscape of Few-Shot Learning. These models, pretrained on internet-scale datasets using self-supervised or contrastive learning objectives, acquire highly generalizable representations that can be adapted to novel tasks with minimal or no task-specific training. This capability, often termed emergent few-shot performance, challenges the traditional boundaries of FSL research.

Foundation models approach the few-shot problem through a different mechanism than classical meta-learning. Rather than learning an explicit adaptation algorithm through episodic training, these models leverage the vast diversity of their pretraining data to implicitly encode a broad spectrum of

visual, linguistic, and conceptual relationships. When presented with a novel task defined by a few examples, the model can leverage these pre-encoded relationships to make accurate predictions without gradient-based fine-tuning. For instance, CLIP's joint embedding of images and text enables zero-shot and few-shot classification through natural language prompts, bypassing the need for task-specific architectures (Radford et al., 2021).

The implications of this paradigm shift for FSL are profound. First, the performance gap between foundation models and specialized meta-learning methods has narrowed considerably, with models like GPT-4 achieving competitive few-shot results across diverse domains without domain-specific architectural engineering. Second, the distinction between pretraining and adaptation is becoming increasingly blurred, as foundation models can be prompted or conditioned on task descriptions rather than requiring explicit support sets.

However, foundation models also introduce new challenges for FSL. Their massive scale makes them computationally prohibitive for resource-constrained environments, contradicting one of the original motivations for FSL—efficient learning. Furthermore, their generalization capabilities, while impressive, are not guaranteed and can fail systematically on out-of-distribution tasks or domains underrepresented in pretraining data. The black-box nature of these models also complicates the application of the security analyses presented in Section 5, as adversarial vulnerabilities may exist in latent spaces that are neither interpretable nor directly accessible.

Despite these challenges, the trajectory of FSL research is increasingly converging with foundation model development. Future FSL systems will likely adopt hybrid architectures that combine the efficiency and interpretability of classical meta-learning with the representational power of foundation models, achieving the best of both paradigms.

7.3. In-Context Learning as Emergent Few-Shot Capability

A particularly significant development arising from foundation models is in-context learning (ICL), a phenomenon where large language and vision models learn to perform new tasks simply from examples provided within the input context, without any parameter updates. First systematically observed in GPT-3, in-context learning has since been demonstrated across modalities and represents a radical departure from conventional gradient-based adaptation (Brown et al., 2020).

In the context of Few-Shot Learning, in-context learning can be understood as an extreme form of few-shot adaptation where the learning occurs entirely

at inference time through attention mechanisms. The model does not update its weights; instead, it reconfigures its internal computation based on the contextual relationships among the provided examples and the query. This mechanism bears a conceptual resemblance to metric-based meta-learning, where classification is performed by comparing the query to support examples in an embedding space. However, in-context learning operates in a vastly higher-dimensional and more flexible representational space, enabled by the scale of the underlying model.

The relationship between in-context learning and classical FSL raises important theoretical questions. Research has shown that transformer-based in-context learning can implement gradient descent algorithmically within its forward pass, effectively simulating the adaptation process of optimization-based meta-learners without explicit parameter updates. This suggests a deep structural connection between the two paradigms, with in-context learning representing a more implicit and scalable realization of meta-learning principles (von Oswald et al., 2023).

For practical FSL applications, in-context learning offers several advantages. It eliminates the need for episodic training and task design, reducing the engineering overhead associated with classical meta-learning. It also enables seamless integration of multimodal information, as contextual examples can include text descriptions, images, and structured data simultaneously. However, the effectiveness of in-context learning is highly sensitive to prompt design—the selection, ordering, and formatting of examples significantly influence performance, a phenomenon known as prompt sensitivity or prompt brittleness.

Moreover, in-context learning inherits the security vulnerabilities discussed in Section 5, albeit in modified forms. Adversarial perturbations can be applied to contextual examples to manipulate model predictions, representing a new variant of feature-level attack. Poisoning attacks can target the examples retrieved from external knowledge bases to populate the context, and membership inference risks persist regarding whether specific examples were included in the model's pretraining data.

Looking forward, the integration of in-context learning with classical FSL methodologies represents a promising research direction. Hybrid approaches that use meta-learning to optimize prompt templates or example selection strategies for in-context learning could combine the efficiency of explicit adaptation with the representational power of large foundation models. Such integration would mark a significant step toward truly human-like few-shot learning systems that can rapidly acquire new concepts from minimal examples while maintaining robustness, interpretability, and security.

References

Cao, T., Law, M. T., & Fidler, S. (2020). *A theoretical analysis of the number of shots in few-shot learning*. arXiv preprint arXiv:1909.11722.

Da Silva, F. L., & Costa, A. H. R. (2019). *A survey on transfer learning for multiagent reinforcement learning systems*. *Journal of Artificial Intelligence Research, 64*, 645–703.

Finn, C., Abbeel, P., & Levine, S. (2017). *Model-agnostic meta-learning for fast adaptation of deep networks*. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. *International Conference on Learning Representations (ICLR)*.

Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Kim, G. N., Lee, H. J., Jeong, I. W., Shin, J. M., & Choi, S. H. (2026). *EAMF: Feature-level adversarial attack on metric-based few-shot learning models*. *IEEE Access*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. *International Conference on Learning Representations (ICLR)*.

Munkhdalai, T., & Yu, H. (2017). *Meta networks*. In *International Conference on Machine Learning*.

Parnami, A., & Lee, M. (2021). *Learning from few examples: A summary of approaches to few-shot learning*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Payandeh, A., Baghaei, K. T., Fayyazsanavi, P., Ramezani, S. B., Chen, Z., & Rahimi, S. (2023). *Deep representation learning: Fundamentals, technologies, applications, and open challenges*. *IEEE Access, 11*, 137621–137659.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., ... & Wu, Q. J. (2022). *A review of generalized zero-shot learning methods*. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4051–4070.

Schwartz, E., Karlinsky, L., Feris, R., Giryes, R., & Bronstein, A. (2022). *Baby steps towards few-shot learning with multiple semantics*. *Pattern Recognition Letters, 160*, 142–147.

Shu, J., Xu, Z., & Meng, D. (2018). *Small sample learning in big data era*. arXiv preprint arXiv:1808.04572.

Snell, J., Swersky, K., & Zemel, R. (2017). *Prototypical networks for few-shot learning*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). *A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities*. *ACM Computing Surveys, 55*(13s), 1–40.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). *Learning to compare: Relation network for few-shot learning*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). *Matching networks for one shot learning*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xian, Y., Korbar, B., Douze, M., Schiele, B., Akata, Z., & Torresani, L. (2020). *Generalized many-way few-shot video classification*. In *European Conference on Computer Vision* (pp. 111–127). Springer International Publishing.

Xu, X., Kamath, S., Butt, M. A., & Raducanu, B. (2025, October). *An h-space based adversarial attack for protection against few-shot personalization*. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 4904–4913).

Zhang, C., Hu, M., Li, W., & Wang, L. (2025). *Adversarial attacks and defenses on text-to-image diffusion models: A survey*. *Information Fusion, 114*, 102701, 1–15.

Zhao, J., Kong, L., & Lv, J. (2025). *An overview of deep neural networks for few-shot learning*. *Big Data Mining and Analytics, 8*(1), 145–188.

Zheng, B., Liang, C., & Wu, X. (2025). *Targeted attack improves protection against unauthorized diffusion customization*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901.

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, 47230–47244.

Liu, Y., Ma, X., Bailey, J., & Lu, F. (2020). Reflection backdoor: A natural backdoor attack on deep neural networks. In European Conference on Computer Vision (ECCV) (pp. 182–199).

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML) (pp. 8748–8763).

Shafahi, A., Huang, W. R., Najibi, M., et al. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 31.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In IEEE Symposium on Security and Privacy (SP) (pp. 3–18).

Tramer, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In USENIX Security Symposium (pp. 601–618).

von Oswald, J., Niklasson, E., Randazzo, E., et al. (2023). Transformers learn in-context by gradient descent. In International Conference on Machine Learning (ICML) (pp. 35151–35174).

Bölüm 4

Büyük Dil Modellerinin (LLM) Kuramsal Sınırları ve Varsayımları

Tevfik Erdal Baylav[1]

Atınç Yılmaz[2]

Özet

Bu bölüm, büyük dil modellerinin kuramsal sınırlarını ve varsayımlarını, sistem başarısızlığı göstergesi değil, kullanım bağlamının sınırlarını tanımlayan özellikler olarak ele alır. Benchmark skorlarının anlamı hangi varsayımların devrede olduğuna bağlıdır; sınır analizi bu varsayım kümesini ve dolayısıyla karar veya bilgi kaynağı olarak modelin hangi bağlamlarda konumlandırılabileceğini görünür kılar. Dilin olasılıksal temsili, sonlu bağlam penceresi ve token düzeyinde eğitim hedefi hesaplama ve temsil varsayımlarını oluşturur; genelleme eğitim–değerlendirme dağılımları ve görev formatına bağımlıdır. Çıktının epistemik güvenilirliği bağlama bağımlıdır; dağıtım koşulları ve doğrulama pratikleri tasarım varsayımlarıyla örtüştüğünde çıktı bilgi kaynağı işlevi görebilir, aksi durumda güvenilirlik ve belirsizlik uygulama katmanında yönetilir. Halüsinasyon tasarım özelliği olarak, hangi görevlerde retrieval, doğrulama modülleri veya insan-in-the-loop ile sınırlandırılması gerektiğinin gerekçesini oluşturur. Açıklanabilirlik ve agent mimarileri, sınırların pratikte nasıl yönetileceği ve hangi mimari güvencelerin devreye alınacağı sorusunu gündeme getirir. Bölüm, LLM etkinliğinin bağlama bağımlı olduğunu ve retrieval, insan denetimi ve modüler doğrulama katmanlarıyla artırılabileceğini vurgulayarak sonuçlandırır.

1 MSc, PhD Candidate, Department of Computer Engineering, Beykent University, Istanbul, Turkey, baylav.tevfik@gmail.com

2 Assoc. Prof. Dr., Faculty of Applied Sciences, Marmara University, Istanbul, Turkey, atinc.yilmaz@marmara.edu.tr

https://doi.org/10.58830/ozgur.pub1351.c5537

1. Kuramsal Çerçeve ve Kullanım Çerçevesi

Benchmark skorları anlam taşıyor—taşıdıkları anlam hangi varsayımların devrede olduğuna bağlı. LLM'lerin sınırlarına bakmak, sistemin başarısızlığını değil, kullanım bağlamının sınırlarını tanımlayan özellikleri öne çıkarır. Tasarımın getirdiği bu özellikler, hangi bağlamlarda modelin karar veya bilgi kaynağı olarak konumlandırılabileceğini belirler; dolayısıyla sınır analizi, epistemolojik bir çerçeve sunarak uygulama katmanında dağıtım ve tasarım kararlarının temelini oluşturur.

Model, verilen token dizisinden sonraki tokenın olasılığını tahmin eder. Dil burada iletişim aracı veya anlam taşıyıcı değil, ayrık sembollerin ardışık olasılık yapısı olarak temsil edilir; semantik ve pragmatik yalnızca bu yapıya yansıdığı ölçüde modele girer. Bu temsil hesaplanabilirlik için neredeyse zorunludur; "anlama" ile "yüksek olasılıklı çıktı üretme" ise aynı şey değildir (Bender ve ark., 2021). Bu ayrım, olasılıksal çıktının bilgi ve karar sistemlerinde hangi bağlamlarda kaynak olarak kullanılabileceğinin sınırını tanımlar. Chomsky'ci perspektif dilin yalnızca gözlemlenebilir dizilerin istatistiği olmadığını vurgular (Chomsky, 1965); LLM tarafı ise kuralların veri içinde örtük öğrenildiğini varsayar (Brown ve ark., 2020). Bu metinde vurgu, bu temsilin hangi kullanım sınırlarını belirlediği ve bu sınırların sistem tasarımında nasıl dikkate alınacağı üzerinedir.

"İyi çalışma hangi koşullarda tanımlanır?" sorusu, başarı metriklerinin anlamlı olduğu varsayım kümesini netleştirir. Bu varsayımlar ihlal edildiğinde metrikler yanıltıcı hale gelebilir; sınır analizi, tam da bu varsayım kümesini ve dolayısıyla dağıtım bağlamını görünür kılar.

2. Hesaplama ve Temsil Varsayımları

2.1. Bağlam Penceresi ve Mimari Sınırları

Bağlam penceresi sonludur—bu bir arıza değil, mimarinin tanımlayıcı özelliğidir. Pencere dışındaki bilgi model için yapısal olarak erişilemez; temsilin sınırı burada çizilir. Transformer mimarisinde dikkat mekanizması $O(n^2)$ karmaşıklığa sahip olduğundan pencereyi büyütmek maliyet ve bellek getirir; sonsuz pencere de "tüm metni anlamak" değil, daha uzun sonlu dizi işlemek anlamına gelir (Vaswani ve ark., 2017). Bu özellik, modelin hangi soru ve görev türlerinde anlamlı yanıt üretebileceğinin sınırını belirler; sistem tasarımında bağlam gereksinimi ve retrieval-augmented generation (RAG) gibi tamamlayıcı mimariler bu sınırla uyumlu biçimde seçilebilir (Lewis ve ark., 2020).

2.2. Eğitim Hedefi, Ölçek ve Bağlam Bağımlılığı

Hesaplama varsayımları—sonlu parametre, sonlu bağlam, token düzeyinde çapraz entropi hedefi—modelin hangi fonksiyonları temsil edebileceğini yapısal olarak koşullandırır. Eğitim hedefi, T uzunluğundaki bir dizi için negatif log-olabilirlik minimize edilmesi olarak şu biçimde ifade edilebilir:

$$L(\theta) = -(1/T) \Sigma \log P(x_t \mid x_1, \ldots, x_{t-1}; \theta)$$

Bu hedef, modelin "en yüksek olasılıklı temsil"i öğrenmesini sağlar—"tüm geçerli doğrular" veya "bağlama göre en uygun" yanıtı değil. Çoklu geçerli yorumlar veya belirsizlik altında stratejik tercihler eğitim hedefine dolaylı yansır. Eğitim hedefi ile kullanım bağlamı arasındaki bu ilişki, hangi bağlamlarda modelin çıktısının doğrudan kullanılabileceğini, hangi bağlamlarda ise insan denetimi veya modüler doğrulama katmanlarıyla desteklenmesi gerektiğini belirler. Ölçek yasaları belirli metriklerde iyileşme vaat eder (Kaplan ve ark., 2020; Hoffmann ve ark., 2022); bu metrikler genelleme, tutarlılık veya güvenilirlikle özdeş değildir. Sınırlar sabit değil, kullanım bağlamına göre değişkendir—etkinlik bağlama bağımlıdır ve mimari seçimlerle kullanım bağlamı içinde kalınarak artırılabilir.

2.3. Formal Problem Statement: LLM as Bounded Stochastic Sequence Estimator

LLM, mühendislik perspektifinden sınırlı bir stokastik dizi tahmincisi olarak tanımlanabilir. Model, bir vokabüler V üzerinde tanımlı token dizilerinin koşullu olasılık dağılımını parametrize eder:

$$P\_\theta(x_1, \ldots, x_n) = \Pi\ P\_\theta(x_t \mid x_1, \ldots, x_{t-1})$$

Burada θ, eğitim verisi D üzerinden ampirik risk minimizasyonuyla öğrenilen parametre kümesidir. Model P_θ, gerçek veri dağılımı P_data'nın bir yaklaşımıdır; bu yaklaşımın kalitesi dağılım kayması (distribution shift) durumunda garanti edilemez. Bu formal çerçeve, modelin ne yaptığını—dağılıma uygun token dizisi üretmek—ve ne yapmadığını—gerçeği doğrulamak, nedensel çıkarım yapmak—açıkça tanımlar. Dağıtım kararları bu sınırlılık çerçevesinde alınmalıdır.

3. Genelleme ve Dağılımsal Sınırlar

3.1. Dağılım ve Görev Bağımlılığı

Genelleme, eğitim ve değerlendirme dağılımları arasındaki ilişkiye bağlıdır; bu ilişki görev-uyumlu kullanım alanının tanımında merkezi rol oynar. Eğitim dağılımı dışındaki girdilere çıktı kalitesi, test dağılımının eğitime yakınlığına

bağlıdır; görev genellemesi ise modelin belirli formatlarda eğitilip ince ayarlanması nedeniyle format ve görev tanımına duyarlıdır. Tıp alanında bu mesele somut biçimde gündeme gelmiştir: klinisyenlerin LLM çıktısını doğrudan klinik karar verme sürecine dahil ettiği durumlarda, modelin dağılım dışı sorgulara verdiği yanıtların güvenilirlik sorunları gözlemlenmiştir (Singhal ve ark., 2023). Bu özellikler başarısızlık göstergesi değil, dağıtım kararlarının dayandığı tasarım bilgisidir.

3.2. Dünya Bilgisi, Zamansal Sürüklenme ve Format Duyarlılığı

"Dünya bilgisi" örtük temsilde kodlanır; açık ontoloji veya nedensel grafik yoktur. Aynı bilgi farklı ifadeyle sorulduğunda tutarsızlık veya güncelliğini yitirmiş bilginin sunulması bu temsilin doğal sonucudur. Güncelleme etkisi izlenebilir değildir; ince ayar veya müdahale yan etkileri öngörülemez (Zhu ve ark., 2020). Bu özellikler, modelin hangi bilgi türleri ve zaman dilimleri için uygun olduğunu, hangi durumlarda retrieval veya dış veri kaynaklarıyla desteklenmesi gerektiğini belirler. Zamansal sürüklenme ve format duyarlılığı, genellemenin sunum ve bağlama bağımlı olduğunu gösterir; bu da sistem tasarımında bağlam eşlemesi ve periyodik güncelleme stratejilerinin gerekçesidir.

4. Epistemolojik Sınırlar ve Bilgi Statüsü

4.1. Olasılıksal Çıktı, Atıf ve İçerik-Form Ayrımı

Model doğruluk iddiasında bulunmaz; yalnızca veri dağılımına uygunluk gösterir. Kullanıcı çıktıyı "bilgi" olarak aldığında bu bir atıftır—model bu atfı doğrulayacak mekanizma sunmaz. Epistemoloji literatüründe bilgi için genellikle üç koşul aranır: doğruluk (truth), gerekçelendirilmiş inanç (justified belief) ve güvenilirlik (reliability) (Goldman, 1979). LLM çıktısı bu koşulların hiçbirini içsel olarak karşılamaz; çıktının bilgi statüsü kazanması, dış doğrulama mekanizmalarına bağımlıdır. İçerik ile yüzey formu ayrımı da bu bağlamda önemlidir: model anlamsal içeriği doğrudan işlemez, token dizilerinin olasılık yapısını işler.

Aşağıdaki tablo epistemolojik katmanlar ile bunlara karşılık gelen sistem düzeyi telafi mekanizmalarını özetlemektedir:

Tablo 1. Epistemolojik Katmanlar ve Sistem Düzeyi Telafi Mekanizmaları

Katman	Model İçsel Özellik	Sistem Düzeyi Telafi
Olasılık	Cross-entropy eğitim hedefi	Retrieval mekanizmaları (RAG)
Belirsizlik	Aleatorik/epistemik ayrımı yapılamaz	Kalibrasyon katmanı (temperature scaling)
Halüsinasyon	Distributional fit, gerçeklik değil	Harici doğrulayıcı / insan denetimi
Açıklanabilirlik	Nedensel zincir kurulamaz	Modüler doğrulama, yapısal çıktı

4.2. Belirsizlik, Gerekçelendirme ve Açıklanabilirlik Tasarımı

Olasılık dağılımı aleatorik ile epistemik belirsizliği ayırmaz; çıktı olasılıkları "bilmiyorum" sinyali olarak güvenilir biçimde kullanılamaz (Kadavath ve ark., 2022). Kalibrasyon araştırmaları, modelin token başına ürettiği olasılık değerlerinin gerçek doğruluk olasılığına ne ölçüde karşılık geldiğini inceler; bu değerlerin doğrudan karar girişi olarak kullanılması yanıltıcı olabilir (Guo ve ark., 2017). Bu özellikler, otonom veya karar-destek ortamlarında LLM çıktısının nasıl konumlandırılacağını—ham veri mi, yoksa insan veya üst sistem tarafından olasılık atamasıyla birleştirilecek girdi mi—belirleyen tasarım ölçütleridir. Gerekçelendirme ve chain-of-thought çıktısının nedensel ya da mantıksal rolü belirsizdir; tutarlılık doğruluk için gerekli ama yeterli değildir (Wei ve ark., 2022). Bu da açıklanabilirlik ve hesap verebilirlik gereksinimlerinin uygulama katmanında nasıl karşılanacağının tasarım konusu olduğunu gösterir.

5. Karar Desteği ve Otonom Sistemlerde Tasarım

5.1. Halüsinasyon, Görev-Uyumlu Kullanım Alanı ve Mimari Güvenceler

Halüsinasyon, modelin hedefinin "gerçeğe uygunluk" değil "veri dağılımına uygunluk" olmasının doğal sonucudur—tasarım özelliği, arıza değil (Ji ve ark., 2023). Bu özellik, hangi görevlerde LLM çıktısının doğrudan eyleme dönüştürülebileceğini, hangi görevlerde ise retrieval, doğrulama modülleri veya insan-in-the-loop ile sınırlandırılması gerektiğini belirler. Hukuki belge üretiminde bu sınır özellikle kritiktir: LLM'lerin mahkeme içtihatlarında var olmayan atıflar ürettiği belgelenmiştir (Magesh ve ark., 2024). Bu tür vakalar, yüksek riskli alanlarda harici doğrulayıcı katmanının tasarım zorunluluğu olduğunu somutlaştırmaktadır. Otonom sistemlerde model çıktısı eyleme dönüşebilir; otomasyon piramidi—insanın loop'ta, üzerinde veya dışında olması—tasarım seçimidir.

5.2. Tekrarlanabilirlik, Kalibrasyon ve Sistem Düzeyi Etkinlik

Tekrarlanabilirlik ve doğruluk bağlama bağımlıdır; görev-spesifik risk profili ve kabul edilebilir eşikler tasarımın parçası olarak tanımlandığında görev-uyumlu kullanım alanı netleşir. LLM çıktısı olasılık tahminlerini açık sunmuyorsa, karar verici çıktıyı ham veri alıp kendi olasılık atamasıyla birleştirir—bu birleştirme, uygun mimari seçimlerle yönetilir. Retrieval, doğrulama modülleri ve insan-in-the-loop, kuramsal özellikleri değiştirmez; ancak bu özelliklerin tanımladığı sınırlar içinde etkinliği artırır ve kullanım bağlamını genişletir. Otonom araç sistemlerinde de benzer bir yaklaşım gözlemlenmektedir: LLM, algı ve planlama bileşenlerinin çıktısını yorumlamak için kullanılmakta, ancak kritik güvenlik kararları kural tabanlı doğrulama katmanlarıyla denetlenmektedir (Wen ve ark., 2023).

6. Açıklanabilirlik ve Agent Mimarileri

6.1. Açıklama, Doğrulama ve Hesap Verebilirlik

"Neden bu cevap?" sorusu LLM'de nedensel zincir veya kural tabanlı gerekçe ile yanıtlanmaz; karar milyarlarca parametrenin etkileşimidir. Global açıklama pratikte ulaşılamaz; lokal açıklama kısmen sağlanabilir (attention ağırlıkları, integrated gradients vb.) ancak nedensel yanıt vermez (Ribeiro ve ark., 2016; Lipton, 2018). Çıktının formal veya mantıksal doğrulanması birçok karar alanında istenir; doğal dil olduğu için otomatik doğrulama ek bileşen (yapısal forma çevirme, modüler doğrulama katmanı) gerektirir. "Model ne dedi?" yanıtlanabilir; "modelin dediği geçerli mi?" sorusu ise sistem tasarımında insan veya doğrulama modülüyle konumlandırılır. Bu ayrım, tasarım güvencelerinin seçiminde merkezi rol oynar.

6.2. Agent Tasarımı ve Sınırları Sarmalayan Katmanlar

Agent mimarilerinde LLM planlama, araç kullanımı ve çok adımlı görevlerde merkezi bileşen olarak kullanılır (Yao ve ark., 2023; Shinn ve ark., 2023). Uzun zincirlerde tutarlılık ve dış dünya geri bildirimine tepki, tasarımın dikkate aldığı özelliklerdir; hata yayılımı ve araç yan etkilerinin temsili bu özelliklerle sınırlıdır. Dilsel temsil ile dünya durumları arasındaki eşleme model tarafından garanti edilmez; bu eşleme, plan doğrulayıcı, insan denetimi, modüler kontrol gibi tasarım ve doğrulama katmanlarıyla kısmen telafi edilir. Agent mimarisi, LLM'in tanımladığı sınırları sarmalayan katman olarak tasarlanır; LLM sınırsız karar verici olarak değil, görev-uyumlu kullanım alanı içinde konumlandırılarak etkinlik artırılır.

7. Sonuç ve Tasarım Ölçütleri

Belirlenen özellikler tek sonuca indirgenemez; farklı katmanlar farklı tür kullanım sınırları tanımlar. Bu sınırlar aşılacak kısıtlar değil, seçilen temsil ve öğrenme paradigmasının tanımlayıcı sonuçlarıdır; dağıtım bağlamına uygun tasarım kararları bu sınırlara göre alınır. Hesaplama varsayımları modelin neyi temsil edebileceğini yapısal olarak koşullandırır; genelleme ve dağılımsal özellikler başarının hangi eğitim–test ve görev bağlamlarında geçerli olduğunu belirler; epistemolojik özellikler çıktının bilgi statüsünün atıftan ibaret olduğunu ve belirsizliğin uygulama katmanında nasıl yönetileceğini gösterir.

Otonom karar ve karar-destek ortamlarında güvenilirlik, bu özelliklerin tanımladığı sınırlar içinde bir tasarım meselesidir; halüsinasyon ve tutarsızlık arıza değil, hangi mimari önlemlerin devreye alınacağının gerekçesidir. Açıklanabilirlik ve agent sistemleri, sınırların pratikte nasıl yönetileceği ve hangi bağlamlarda LLM'in uygun bileşen olarak konumlandırılacağı sorusunu gündeme getirir.

"LLM'ler güvenilir mi?" sorusu, "hangi varsayımlar altında, hangi görevlerde, hangi doğrulama ve denetim katmanlarıyla görev-uyumlu kullanım alanı içinde etkindir?" biçiminde çerçevelendiğinde anlamlı yanıt verir. Teknik özellikler yasaklama veya sınırsız kabul ikilemine indirgenemez; sınırların açık ifadesi, risk tabanlı düzenleme ve sorumlu dağıtım için ön koşuldur ve bağlama bağımlı etkinliğin retrieval, insan denetimi ve modüler doğrulama gibi mimari önlemlerle nasıl desteklenebileceğinin temelini oluşturur.

Kaynakça

Bender, E. M., Gebru, T., McMillan-Major, A., ve Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT 2021, 610–623.

Brown, T., Mann, B., Ryder, N., ve ark. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), Justification and Knowledge. Reidel.

Guo, C., Pleiss, G., Sun, Y., ve Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of ICML 2017, 1321–1330.

Hoffmann, J., Borgeaud, S., Mensch, A., ve ark. (2022). Training compute-optimal large language models. arXiv:2203.15556.

Ji, Z., Lee, N., Frieske, R., ve ark. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Kadavath, S., Conerly, T., Askell, A., ve ark. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Kaplan, J., McCandlish, S., Henighan, T., ve ark. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Lewis, P., Perez, E., Piktus, A., ve ark. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

Magesh, V., Surani, F., Dahl, M., ve ark. (2024). Hallucination-free? Assessing the reliability of leading AI legal research tools. arXiv:2405.20362.

Ribeiro, M. T., Singh, S., ve Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of KDD 2016, 1135–1144.

Shinn, N., Cassano, F., Berman, E., ve ark. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv:2303.11366.

Singhal, K., Azizi, S., Tu, T., ve ark. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180.

Vaswani, A., Shazeer, N., Parmar, N., ve ark. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wei, J., Wang, X., Schuurmans, D., ve ark. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35.

Wen, L., Yang, X., Fu, D., ve ark. (2023). Road GPT: Unifying large language model and autonomous driving. arXiv:2311.10813.

Yao, S., Zhao, J., Yu, D., ve ark. (2023). ReAct: Synergizing reasoning and acting in language models. Proceedings of ICLR 2023.

Zhu, C., Chen, A., Shen, T., ve ark. (2020). Modifying memories in transformer models. arXiv:2012.00363.

Chapter 5

The Role of Artificial Intelligence and Big Data in Transforming Modern Cybersecurity

Sara Naghib Zadeh[1]

Cansu Arslan[2]

Abstract

In the era of digital transformation, cybersecurity has evolved from a technical necessity into a fundamental pillar for organizational resilience. The rapid proliferation of cloud computing, the Internet of Things (IoT), and integrated enterprise ecosystems has streamlined operations but simultaneously expanded the cyber-attack surface (Saeed et al., 2023). Modern adversaries now leverage automated and intelligent techniques, rendering traditional, rule-based security measures insufficient against complex threats such as ransomware, supply chain compromises, and zero-day breaches (Lahare & Wakchaure, 2025). Within this landscape, Enterprise Resource Planning (ERP) systems represent a critical strategic asset. As the operational backbone of organizations, they manage highly sensitive data across finance, human resources, and customer relations. The migration of ERP systems to cloud environments and their integration with diverse APIs has introduced new vulnerabilities, including misconfigurations and sophisticated phishing attacks. Consequently, protecting these high-value targets in a hyper-connected world requires a shift from reactive monitoring to proactive defense(Bhat & Jayaram, 2025). The exponential growth of security-related data, encompassing system logs, network traffic, and user behavior, has turned modern cybersecurity into a "Big Data" challenge. Analyzing such vast and heterogeneous datasets in real-time is beyond human capacity and traditional signature-based tools. To address this, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as essential solutions, offering the ability to detect hidden patterns, identify anomalies, and respond to unknown threats autonomously (Mohamed, 2025). This study provides a comprehensive

1 Dr. Lecture, Halic University, Vocational School, Department of Computer Programming, ORCID: 0009-0005-6959-1165.

2 Halic University, Vocational School, Department of Big Data Analytics, ORCID:0009-0006-7532-5418

https://doi.org/10.58830/ozgur.pub1351.c5538

analysis of how Big Data analytics and AI-driven models are transforming cybersecurity within cloud-based ERP environments. By examining the synergy between intelligent algorithms and large-scale data infrastructures, the paper identifies key implementation challenges and proposes a strategic foundation for next-generation defense systems.

1. Introduction

In the era of digital transformation, cybersecurity has evolved from a technical necessity into a fundamental pillar for organizational resilience. The rapid proliferation of cloud computing, the Internet of Things (IoT), and integrated enterprise ecosystems has streamlined operations but simultaneously expanded the cyber-attack surface (Saeed et al., 2023). Modern adversaries now leverage automated and intelligent techniques, rendering traditional, rule-based security measures insufficient against complex threats such as ransomware, supply chain compromises, and zero-day breaches (Lahare & Wakchaure, 2025).

Within this landscape, Enterprise Resource Planning (ERP) systems represent a critical strategic asset. As the operational backbone of organizations, they manage highly sensitive data across finance, human resources, and customer relations. The migration of ERP systems to cloud environments and their integration with diverse APIs has introduced new vulnerabilities, including misconfigurations and sophisticated phishing attacks. Consequently, protecting these high-value targets in a hyper-connected world requires a shift from reactive monitoring to proactive defense(Bhat & Jayaram, 2025).

The exponential growth of security-related data, encompassing system logs, network traffic, and user behavior, has turned modern cybersecurity into a "Big Data" challenge. Analyzing such vast and heterogeneous datasets in real-time is beyond human capacity and traditional signature-based tools. To address this, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as essential solutions, offering the ability to detect hidden patterns, identify anomalies, and respond to unknown threats autonomously (Mohamed, 2025).

This study provides a comprehensive analysis of how Big Data analytics and AI-driven models are transforming cybersecurity within cloud-based ERP environments. By examining the synergy between intelligent algorithms and large-scale data infrastructures, the paper identifies key implementation challenges and proposes a strategic foundation for next-generation defense systems.

2. The Need for Transformation in Protecting Critical Infrastructures

Today, critical infrastructures are facing a new paradigm of threats that go beyond traditional intrusion patterns. Incidents such as the disruption of Ukraine's power distribution network and the paralysis of the Colonial Pipeline served as serious warnings for national security worldwide, demonstrating that a single security breach in such systems can trigger cascading and catastrophic consequences for both the economy and public welfare. These real-world events clearly indicate that conventional approaches based on human monitoring and reactive responses are no longer sufficient to counter modern, high-speed cyberattacks(AlArfaj & AlShuaibi, 2025).

While defenders continue to rely on reactive security protocols, attackers are increasingly adopting automated tools and malicious artificial intelligence. These emerging threats exhibit a high degree of adaptability, enabling them to dynamically bypass defensive barriers. In this context, the transition from traditional security paradigms to intelligent and real-time defense is not merely an option but a strategic necessity. Artificial intelligence plays a central role in this transition by analyzing massive volumes of data and network traffic to identify anomalous patterns before a crisis occurs(Lehto, 2022).

Machine learning and deep learning algorithms offer significant potential for early threat detection. Unlike rule-based systems, these models are capable of learning from historical data and can even respond to previously unseen "zero-day" attacks. However, the deployment of such technologies in operational environments faces substantial technical challenges. Much of the existing infrastructure consists of legacy systems that were not originally designed for integration with modern intelligent networks or for supporting computationally intensive AI models (Reddy et al., 2024).

In addition to hardware limitations, the non-stop nature of critical infrastructures such as power plants leaves no room for security testing or trial-and-error approaches. Furthermore, balancing the need for continuous data flow to train AI models with the requirements of data confidentiality and privacy presents another significant challenge for practitioners. Therefore, deploying an effective threat detection system requires the design of integrated architectures that ensure multi-layered security without disrupting essential public services (Kechagias et al., 2022).

Finally, it is important to recognize that artificial intelligence itself can become a target of attack. While machine learning models are increasingly used in cloud-based defense systems, they remain vulnerable to threats such as

data poisoning and adversarial attacks designed to manipulate model behavior (Gong et al., 2020). This study adopts a comprehensive perspective to examine the current state of real-time detection systems and proposes practical solutions for implementing and maintaining intelligent security, ensuring that defensive tools do not themselves become the system's weakest link.

3. Strategic Transformation: From Traditional Defense to AI-Driven Proactive Security

In the modern cybersecurity landscape, artificial intelligence (AI) and machine learning (ML) have evolved from supportive tools into integral components of defense strategies. A key advantage of these technologies over traditional approaches lies in their ability to overcome the limitations of signature-based detection and shift toward predictive modeling. As illustrated in Table 1, the intelligent paradigm, unlike conventional reactive methods, is fundamentally based on a proactive approach. By analyzing complex correlations within large-scale datasets, these models can identify early indicators of attacks before significant damage occurs, a capability reflected in the table under the concept of behavior-based detection (Ahsan et al., 2022).

Extensive research in this field has demonstrated the operational effectiveness of machine learning across three critical domains: intrusion detection systems (IDS), intelligent malware classification, and threat intelligence analysis. In this context, supervised learning algorithms leverage historical labeled data to ensure accurate classification of new data points, while unsupervised learning serves as a safeguard against emerging and previously unseen threats (Ahmed Salman et al., 2023). Furthermore, reinforcement learning has opened new horizons in adaptive response mechanisms, enabling systems to continuously optimize their defense strategies in dynamic threat environments. This capability is highlighted in Table 1 under the "accuracy against unknown attacks" metric, representing one of the key strengths of modern approaches (Ferdous et al., 2023).

Another critical dimension of this transformation is the enhancement of real-time response capabilities and the automation of security processes. AI-based tools can correlate disparate data sources and generate a unified view of the attack chain. This level of automation not only significantly reduces the need for human intervention but also enables response times to reach real-time levels, an essential feature identified in Table 1 as a fundamental distinction between intelligent and traditional systems. Ultimately, the scalability of these technologies positions them as the only viable solution for addressing the growing complexity of cyber threats in the digital era (Jada et al., 2024).

However, deploying these intelligent models in dynamic cyber environments requires addressing core learning theory challenges, most notably **concept drift**. Because cyber threats evolve continuously, the statistical properties of the target variables change over time, rendering static historical training data obsolete. To counter this, next-generation defense systems must transition toward **online learning** paradigms, where models process data streams continuously and update their parameters in real time without requiring full retraining cycles. Furthermore, the defensive perimeter must be fortified against **adversarial learning** tactics. Sophisticated attackers increasingly employ adversarial perturbations to craft evasion attacks or execute data poisoning during the training phase, making the mathematical robustness and verifiability of AI algorithms a critical priority in modern security design.

Table 1: Comparative Analysis of Traditional Security Paradigm vs. AI-Based Intelligent Security

Comparison Metric	Traditional Methods (Signature-Based)	Modern Methods (AI-Based)
Type of Approach	Reactive	Proactive
Detection Basis	Known patterns and signatures	Behavioral analysis and anomaly prediction
Response Speed	Dependent on human analysis and signature updates	Real-time and automated
Scalability	Limited under large-scale data volumes	Highly scalable and Big Data compatible
Accuracy	Vulnerable to zero-day attacks	Capable of detecting new and complex threats

4. The Role of Big Data in Enhancing Cybersecurity and Combating Emerging Threats

In the contemporary cybersecurity landscape, Big Data has become one of the fundamental components for identifying, analyzing, and responding to complex threats. The rapid expansion of the Internet of Things (IoT), cloud platforms, enterprise networks, and smart industries has generated enormous volumes of structured and unstructured data, the analysis of which exceeds the capabilities of traditional security methods (Nugroho et al., 2024). Under such conditions, Big Data analytics enables the extraction of valuable knowledge from extensive information sources and allows organizations to identify potential threats before they escalate into crises. As shown in Table 1, the primary advantage of Big Data in cybersecurity lies in the transition

from limited and reactive analysis toward real-time, intelligent, and proactive monitoring (Kumar Bhardwaj et al., 2024).

The implementation of this data-driven approach requires the integration of diverse information sources. As illustrated in Figure 1, Big Data-based security analytics architectures are generally built upon three primary data sources: system logs, network traffic, and user behavior. System logs provide detailed information regarding events, user activities, and authentication processes, making them highly valuable for detecting unauthorized access attempts. At the network level, the analysis of data flows, traffic packets, and communication patterns can reveal intrusion attempts, Distributed Denial-of-Service (DDoS) attacks, and data exfiltration activities. In addition, user behavior analytics plays a significant role in detecting insider threats, account misuse, and suspicious activities by identifying deviations from normal behavioral patterns (Muhati et al., 2024).

One of the most significant advantages of Big Data is its ability to process massive volumes of information in real time through technologies such as distributed computing, cloud computing, and intelligent analytical frameworks. These infrastructures make it possible to examine millions of security events simultaneously and identify hidden relationships among seemingly unrelated incidents. As a result, Time to Detect (TTD) and Time to Respond (TTR) are significantly reduced—an essential factor in sensitive and mission-critical environments (Adams & Heard, 2016).

However, the true value of Big Data becomes evident when it is integrated with artificial intelligence and machine learning. Intelligent models can learn attack patterns from large-scale datasets, detect anomalous behaviors, and even predict unknown threats or zero-day attacks. This synergy between Big Data and AI has transformed security systems from passive mechanisms into adaptive and automated defense architectures (Ahmad et al., 2023).

Beyond monitoring and detection, Big Data analytics also plays a crucial role in response processes, recovery strategies, and risk management. In complex ecosystems such as Enterprise Resource Planning (ERP) environments, where data is distributed across multiple organizational units, this technology can provide a comprehensive view of the security posture and enable faster and more accurate decision-making. Overall, Big Data is not merely a tool for information storage and management; rather, it represents the foundation of next-generation intelligent cyber defense systems capable of significantly improving resilience against emerging and sophisticated threats (Nugroho et al., 2024).

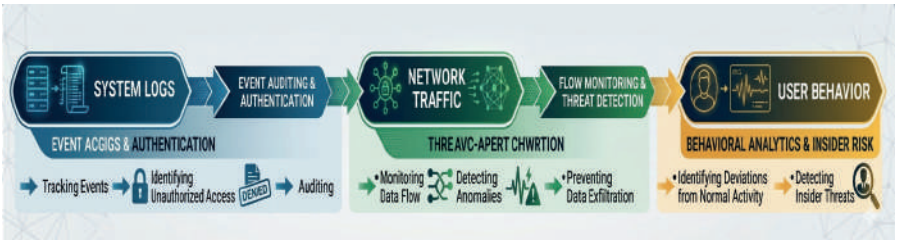


Figure 1. Big Data–Driven Cybersecurity Data Sources: System Logs, Network Traffic, and User Behavior Analytics

5. Exploring Key Challenges in Big Data Analytics for Cybersecurity

Despite its numerous advantages, the adoption of Big Data in cybersecurity is associated with several structural and operational challenges that hinder the full utilization of its potential. Based on the data presented in Figure 2, the most significant obstacle is related to the volume and velocity of data generation, accounting for 30.0% of the overall challenges. This issue creates difficulties in real-time analytics and increases false positive rates, ultimately leading to slower response times and additional pressure on infrastructure resources (Iglesias et al., 2020).

The next major challenges include the scalability of analytical tools and data integration, each representing 20.0% of the identified barriers. Scalability directly affects infrastructure costs and intensifies the need for advanced computational capabilities. Meanwhile, data integration faces problems such as heterogeneous data formats and complex data flows, which complicate the process of aggregating information from multiple sources. In addition, data quality and consistency, accounting for 15.0% of the challenges, emphasize the importance of data cleansing and standardization in achieving accurate analytical outcomes (Alshaibi et al., 2022)

Finally, security and human-related dimensions also constitute an important part of these challenges. Big Data security and privacy concerns (10.0%) involve issues such as encryption gaps and privacy-preserving limitations. Furthermore, the skills gap in data science and cybersecurity, representing 5.0% of the total, highlights the shortage of qualified professionals capable of managing and analyzing such large and complex datasets (Ebunoluwa Johnson et al., 2024).

Overall, these factors indicate that the successful deployment of Big Data-driven cybersecurity systems requires organizations not only to invest in

technical infrastructures, but also to improve data quality and strengthen specialized human expertise.

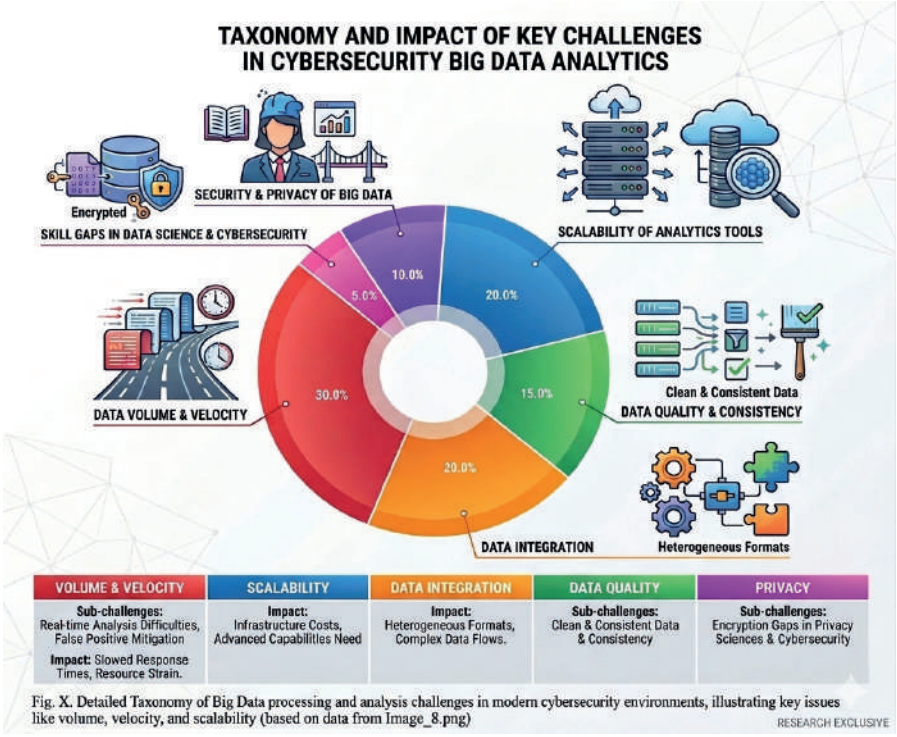


Figure 2. Taxonomy and impact analysis of key challenges in cybersecurity big data analytics. Methodological Note: The taxonomical percentages and impact distribution reported above were derived through a systematic content analysis of the reviewed literature (2020–2026). A thematic coding approach was implemented, where structural and operational barriers identified across 40 peer-reviewed study datasets were categorized into six primary challenge domains. The relative weight of each challenge represents its frequency of occurrence and prioritized impact index within the surveyed research corpus.

6. Intelligent Defense Mechanisms: The Synergy of Machine Learning and Deep Learning

The modern cybersecurity landscape demands a fundamental shift from traditional reactive protocols to intelligent, proactive defense strategies. Traditional methods, primarily reliant on signature-based detection and fixed rules, are increasingly ineffective against the velocity and sophistication of modern threats. As organizations transition to AI-driven security, the primary advantage lies in the ability to overcome the limitations of manual monitoring

and shift toward predictive modeling. The core of this transformation is the capacity of artificial intelligence to analyze complex correlations within massive, heterogeneous datasets, such as system logs, network traffic, and user behavior, to identify early indicators of attacks before significant damage occurs (Kilincer et al., 2021).

The operational process of these intelligent systems begins with a rigorous data ingestion and processing phase. Raw data gathered from multiple sources (IoT devices, cloud logs, and endpoint telemetry) are normalized, cleaned, and prepared for analysis. Subsequently, statistical, temporal, and behavioral features are extracted to be fed into specialized models. This data representation can take various forms, including time series, event sequences, communication graphs, or even binary and image-based formats, enabling the analysis of a wide range of cyberattack scenarios (Okoli et al., 2024).

Machine Learning (ML) serves as the foundational layer of this intelligent defense. In the realm of Supervised Learning, models are trained using labeled data to distinguish between benign and malicious behaviors. Algorithms such as Decision Trees and Random Forests are widely utilized for network traffic classification, while Support Vector Machines (SVM) provide strong performance in intrusion detection based on known patterns. Additionally, models like Naïve Bayes and Logistic Regression remain highly effective for detecting phishing and fraudulent communications. In contrast, Unsupervised Learning becomes essential when data is unlabeled or threats are unknown. Clustering algorithms such as K-Means and DBSCAN group similar behaviors to reveal anomalous patterns, while Principal Component Analysis (PCA) and Isolation Forest are applied to detect outliers and zero-day threats(Chivukula et al., 2023).

At a more advanced level, Deep Learning (DL) architectures extract complex and hidden patterns directly from raw data without the need for manual feature engineering. Convolutional Neural Networks (CNNs) are highly effective for feature extraction in traffic pattern classification and malware detection. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models are specifically suited for sequential and time-series data, making them ideal for analyzing attack sequences and identifying ransomware or Advanced Persistent Threats (APTs). Furthermore, Autoencoders provide superior performance in anomaly detection by identifying unusual deviations from normal system states. To enhance robustness, Generative Adversarial Networks (GANs) can generate synthetic attack scenarios, helping defensive systems "rehearse" against potential evasion techniques (Kohar et al., 2026).

Finally, Natural Language Processing (NLP) serves as a critical component for modern threat intelligence. By utilizing Transformer-based models, security systems can analyze unstructured text from security reports, blogs, and dark web forums to extract indicators of compromise (IoCs), such as malicious IP addresses and attack techniques. The integration of these diverse AI models, from classical ML to cognitive NLP, enables a multi-layered, automated response mechanism capable of real-time actions like IP blocking, host isolation, and access restriction (Albahri et al., 2025).A comprehensive taxonomy detailing the algorithms used in each of these layers, from supervised learning to adversarial models, is provided in Table 2. While challenges such as computational costs, data quality, and adversarial attacks remain, the synergy of these technologies constitutes the backbone of next-generation cyber defense.

Table 2. Comprehensive Taxonomy of Intelligent Models in Cybersecurity

Model Type	Key Algorithms	Primary Security Applications
Supervised Learning	Decision Tree, Random Forest, SVM	Malicious traffic classification, Intrusion detection, Phishing detection
Unsupervised Learning	K-Means, DBSCAN, Isolation Forest	Behavioral anomaly clustering, Zero-day attack detection
Deep Learning (Spatial)	CNN	Malware detection, Feature extraction from network packets
Deep Learning (Temporal)	RNN, LSTM	Attack sequence analysis, Ransomware detection, APT identification
Dimensionality Reduction / Anomaly Detection	PCA, Autoencoder	Outlier detection, Abnormal behavior identification
Cognitive / Textual (NLP)	Transformers, BERT	Threat intelligence analysis, Social engineering detection
Adversarial / Synthetic Models	GANs	Attack evasion simulation, robustness training of security models

7. Cloud Computing Security Threats and Emerging Challenges

Cloud computing has become a fundamental infrastructure in modern information systems, transforming how organizations access computational resources, storage, and digital services. By shifting from local infrastructure to on-demand internet-based services, cloud computing enables scalable, flexible, and cost-efficient access to IT resources, making it widely adopted across industry, academia, and public sectors(Engineering et al., 2023).

Cloud services are typically delivered through three main models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). These models improve deployment agility, reduce infrastructure costs, and support rapid digital transformation in organizations (Younis et al., 2024).

Despite these advantages, cloud computing introduces significant security challenges due to the centralization of sensitive data and the expansion of the attack surface. Key threats include data breaches, identity theft, and unauthorized access, which can result in severe financial and operational damage. In addition, misconfiguration of cloud resources is a major security risk, often caused by incorrect security settings rather than provider-side failures (Bhushan & Gupta, 2017).

Insider threats and Distributed Denial of Service (DDoS) attacks further complicate cloud security. Authorized users may intentionally or unintentionally leak sensitive data, while DDoS attacks can disrupt service availability. Therefore, cloud security must ensure not only data confidentiality but also system availability and operational continuity(Balani et al., 2020).

Table 3 summarizes the most critical cloud security threats along with their impacts and mitigation strategies. As shown, issues such as data breaches, misconfigurations, insider threats, and identity theft require a combination of encryption, access control, multi-factor authentication, and continuous monitoring to mitigate risks effectively.

Recent advancements in artificial intelligence and machine learning have significantly enhanced cloud security capabilities. These methods enable real-time anomaly detection by learning normal user and system behavior patterns. Abnormal activities such as unusual login locations, sudden traffic spikes, or unauthorized access attempts can be automatically detected and responded to, reducing reaction time and improving incident handling efficiency (Subramanian et al., 2018).

As AI models take on autonomous decision-making roles in cloud security, the 'black-box' nature of deep learning architectures introduces significant trust and verification issues for security analysts. Consequently, the integration of Explainable AI (XAI) methods—such as SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)—has become essential. XAI frameworks provide transparent, interpretable rationales for model predictions, allowing human operators to understand exactly why a specific cloud network packet or user behavior sequence was flagged as anomalous. This accountability drastically reduces false-positive investigation

times and bridges the gap between autonomous AI actions and human-centric Security Operations Center (SOC) workflows.

In conclusion, cloud computing security requires a multi-layered defense strategy integrating technical, organizational, and intelligent approaches. The combination of encryption, access control mechanisms, zero-trust architecture, and AI-driven threat detection is essential for ensuring secure and resilient cloud environments.

Table 3. Major Cloud Computing Security Threats and Countermeasures

Security Threat	Description	Impacts	Proposed Countermeasures
Data Breach	Unauthorized access to data stored in the cloud	Exposure of sensitive information, financial loss	Data encryption, access control
Misconfiguration	Incorrect configuration of cloud resources and services	Data leakage, unintended public exposure	Security auditing, standard configurations
Insider Threats	Misuse of privileges by authorized users	Data theft or deletion	Least privilege principle, user behavior monitoring
DDoS Attacks	Overloading servers with malicious traffic	Service disruption, reduced availability	Cloud firewall, load balancing
Identity Theft	Attacker obtains valid usernames and passwords	Unauthorized system access	Multi-factor authentication (MFA)
Malware and Ransomware	Infection of services and data	Data loss, service compromise	Backup strategies, intelligent anti-malware systems

8. Big Data Analytics and Cybersecurity in ERP Ecosystems

Enterprise Resource Planning (ERP) systems have become one of the core pillars of business process management in modern organizations. These systems integrate functions such as financial management, human resources, supply chain, production, customer relationship management, and organizational performance analytics within a unified platform. Due to the extensive reliance of daily organizational operations on ERP systems, any disruption or security breach can lead to severe financial, operational, and reputational consequences. With the increasing adoption of cloud-based ERP solutions, the attack surface has expanded significantly, making advanced cybersecurity measures more critical than ever (Madhav Jha et al., 2023).

Next-generation ERP systems are accessible through web browsers, mobile applications, APIs, and smart devices. While this level of connectivity enhances efficiency and flexibility, it also introduces additional opportunities for cyber attackers. Threats such as unauthorized access, credential theft, exploitation of web services, and misconfiguration vulnerabilities can compromise the confidentiality and integrity of ERP data. Since these systems store valuable business information and sensitive customer data, they remain highly attractive targets for cybercriminals (P. Chinta et al., 2022).

The conventional architecture of ERP systems typically consists of three main layers: the database layer, the business logic layer, and the presentation layer. The database layer stores the organization's core data; the middle layer executes business processes and rules; and the presentation layer enables user interaction through web interfaces or client applications. Each of these layers can be targeted by attackers. Common threats include malicious code injection at the application layer, exploitation of privileged database access, and vulnerabilities in the host operating system (Olaoye, 2025).

In recent years, ransomware attacks have become one of the most critical threats to ERP ecosystems. In such attacks, adversaries infiltrate systems, encrypt files and databases, and demand a ransom in exchange for restoring access (Efe & Geliş, 2024a). Since ERP systems form the operational backbone of organizations, their disruption can severely impact supply chains, production processes, sales operations, and customer services. In addition to ransomware, phishing attacks remain a major attack vector, as attackers use deceptive emails to trick users into revealing credentials or granting unauthorized access (Raja et al., 2024).

Insider threats also represent a significant challenge in ERP security. Disgruntled employees, careless users, or contractors with legitimate access privileges may intentionally or unintentionally cause data leakage. These threats are often more difficult to detect than external attacks because they originate within the organization's trusted boundaries. Therefore, user behavior monitoring, the principle of least privilege, and comprehensive activity logging are essential security requirements for ERP systems (Omotoye & Chen, 2026).

The large volume of data generated by ERP systems, networks, security logs, peripheral devices, and cloud services has transformed cybersecurity into a big data problem. Traditional security analysis methods are not capable of processing such massive and heterogeneous datasets. In this context, big data analytics combined with artificial intelligence and machine learning can extract anomalous behaviors from millions of events. These technologies enable

the identification of hidden patterns and support rapid threat detection and predictive analysis of future incidents (P. C. R. Chinta et al., 2024).

Deep learning has also gained increasing importance in this domain. Neural network-based models are capable of learning complex relationships between events and detecting threats that cannot be identified using signature-based methods. For instance, analyzing user login sequences, unusual financial transaction patterns, or sudden changes in access privileges can indicate insider attacks or gradual intrusions. The use of these models in Security Operations Centers (SOCs) significantly reduces both threat detection time and response time (Jha, 2022).

Ultimately, security in ERP ecosystems requires a multi-layered and intelligent approach. The implementation of multi-factor authentication, data encryption, security patch management, role-based access control, continuous backup strategies, and intelligent threat monitoring are among the most critical defensive measures.The correlation between specific threats and their corresponding defense measures, such as the use of WAF for web attacks or MFA for phishing, is further detailed in Figure 3. Given the growing dependence of businesses on connected and cloud-based ERP systems, the integration of big data analytics, artificial intelligence, and cybersecurity can provide a more resilient and reliable infrastructure for the future of organizations (Efe & Geliş, 2024b).

Modern enterprise resilience requires an architectural shift toward Zero Trust Architecture (ZTA), operating on the strict principle of 'never trust, always verify.' Within cloud-based environments, ZTA cannot remain static; it must evolve into an AI-native security framework. Recent breakthroughs in Foundation Models and Large Language Models (LLMs) tailored for cybersecurity have enabled the deployment of intelligent Security Copilots. These AI-native architectures ingest heterogeneous data substrates in real time, allowing security teams to query complex log environments using natural language, automate prompt-driven incident playbooks, and synthesize threat intelligence at unprecedented speeds, thereby transforming the speed of enterprise defense from hours to milliseconds.

3.1. Synthesized AI-Big Data Conceptual Framework for Secure Cloud-ERP Ecosystems

- To synthesize the operational synergy of these technologies, this chapter proposes an original conceptual framework that structures enterprise cyber defense into four interconnected functional layers within cloud-ERP ecosystems:

Data Ingestion & Big Data Infrastructure Layer: This foundational layer utilizes distributed frameworks to aggregate high-velocity, heterogeneous data sources simultaneously, including ERP system logs, peripheral API traffic, database query logs, and network telemetry.

Cognitive Analytics & AI Processing Layer: Acting as the computational brain, this layer deploys classical machine learning for baseline traffic classification, alongside deep learning architectures (LSTMs and Autoencoders) for sequential anomaly detection and insider threat identification. Specialized LLMs and Security Copilots operate in parallel to ingest unstructured global threat feeds.

Adaptive Cybersecurity Monitoring Layer: This continuous evaluation layer evaluates processed analytics against a dynamic Zero Trust verification matrix, ensuring real-time behavioral monitoring of identities, endpoints, and micro-segmented ERP network zones.

Automated Response & Orchestration Mechanism: The final layer triggers autonomous mitigation protocols (e.g., immediate host isolation, programmatic API token revocation, and automated web application firewall routing changes) to neutralize verified threats before catastrophic business disruption occurs.

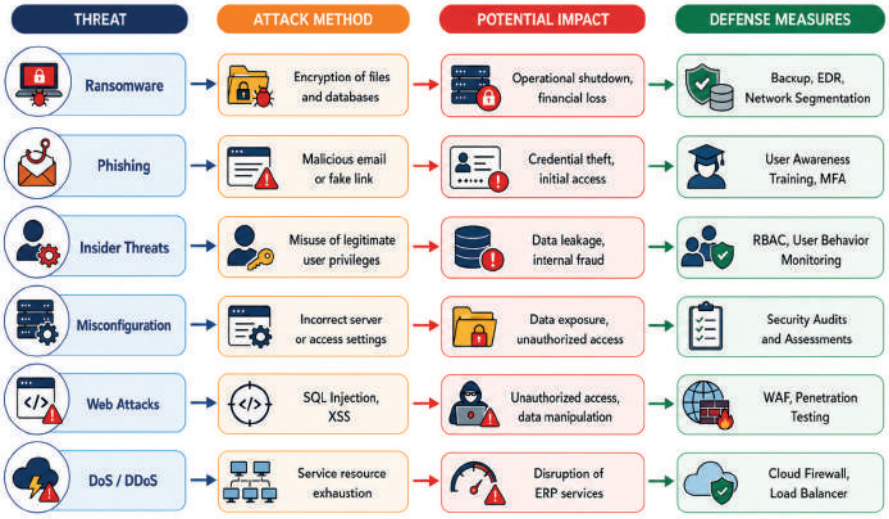


Figure 3. Common Cybersecurity Threats in ERP Ecosystems: Attack Methods, Potential Impacts, and Defense Measures

9. Conclusion

This study provided a comprehensive review of the growing role of artificial intelligence, machine learning, deep learning, and big data analytics in modern cybersecurity, particularly within cloud computing environments and Enterprise Resource Planning (ERP) systems. The results indicate that with the rapid expansion of digital transformation across organizations, the level of cyber threats has significantly increased, making the need for intelligent, adaptive, and scalable security approaches more critical than ever (Kechagias et al., 2022).

Traditional cybersecurity methods, which are mainly based on fixed rules and known signatures, are no longer capable of effectively addressing the complexity and high velocity of modern threats (Shaheen, 2023).In contrast, AI-based approaches enable real-time detection of anomalies and unknown threats by analyzing user and system behavior across large-scale datasets (Khalaf et al., 2025).

It was also observed that the integration of big data analytics with artificial intelligence plays a crucial role in improving threat detection, reducing response time, and increasing accuracy in Security Operations Centers (SOCs) (Jha, 2022). On the other hand, deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and autoencoders have demonstrated strong performance in detecting complex attacks, including advanced persistent threats (APTs) and zero-day attacks (Reddy et al., 2024).

Despite these advantages, several challenges still remain, including data quality issues, high computational costs, adversarial attacks on machine learning models, privacy concerns, and limitations of legacy infrastructures in enterprise systems such as ERP (Efe & Geliş, 2024a).

In conclusion, the future of cybersecurity depends on the development of integrated, intelligent, and multi-layered frameworks that combine big data analytics, artificial intelligence, and zero-trust architecture. Such an approach can provide the necessary resilience, scalability, and robustness against increasingly sophisticated and evolving cyber threats in cloud and enterprise environments.

References

Adams, N., & Heard, N. (2016). Cyber security data sources for dynamic network research. *World ScientificAD KentDynamic Networks and Cyber-Security, 2016•World Scientific*, *1*, 1–211.

Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. (2023). Zero-day attack detection: a systematic literature review. *Artificial Intelligence Review 2023 56:10*, *56*(10), 10733–10811.

Ahmed Salman, H., Alsajri, A., & History, A. (2023). The Evolution of Cybersecurity Threats and Strategies for Effective Protection. A review. *SHIFRA*, *2023*, 73–85.

Ahsan, M., Nygard, K. E., Gomes, R., Chowdhury, M. M., Rifat, N., & Connolly, J. F. (2022). Cybersecurity Threats and Their Mitigation Approaches Using Machine Learning—A Review. *Journal of Cybersecurity and Privacy 2022, Vol. 2, Pages 527-555*, *2*(3), 527–555.

AlArfaj, L., & AlShuaibi, A. (2025). Critical infrastructure protection. *Intelligent and Secure Solutions for Digital Transformation*, 107–130.

Albahri, A. S., Jassim, M. M., Alzubaidi, L., Hamid, R. A., Ahmed, M. A., Al-Qaysi, Z. T., Albahri, O. S., Alamoodi, A. H., Alqaysi, M. E., Mohammed, T. J., Kou, G., Alotaibi, F. S., & Sharaf, I. M. (2025). A trustworthy and explainable framework for benchmarking hybrid deep learning models based on chest X-ray analysis in CAD systems. *World ScientificAS Albahri, MM Jassim, L Alzubaidi, RA Hamid, MA Ahmed, ZT Al-Qaysi, OS AlbahriInternational Journal of Information Technology & Decision Making, 2025•World Scientific*, *24*(8), 2533–2585.

Alshaibi, A., Al-Ani, M., Al-Azzawi, A., Konev, A., & Shelupanov, A. (2022). The Comparison of Cybersecurity Datasets. *Data 2022, Vol. 7, Page 22*, *7*(2), 22.

Balani, Z., & Verma-Salgaokar, H. (2020). Cloud computing security challenges and threats. *International Journal of Computer Science and Mobile Computing*, 9(7), 185–191.

Bhat, J., & Jayaram, Y. (2025). AI-Enhanced Integrations: Secure API Management for Multi-Cloud ERP Environments. *International Journal of Emerging Trends in Computer Science and Information Technology*, *6*(3), 94–103.

Bhushan, K., & Gupta, B. B. (2017). Security challenges in cloud computing: state-of-art. *International Journal of Big Data Intelligence*, *4*(2), 81.

Chinta, P. C. R., Jha, K. M., Velaga, V., Moore, C., Routhu, K., & SADARAM, G. (2024). Harnessing Big Data and AI-Driven ERP Systems to Enhance Cybersecurity Resilience in Real-Time Threat Environments. *SSRN Electronic Journal*.

Chinta, P. C. R., Jha, K. M., Velaga, V., Moore, C., Routhu, K., & Sadaram, G. (2022). AI and ML applications in Big Data analytics: Transforming ERP security models for modern enterprises. *SSRN Electronic Journal*.

Chivukula, A., Yang, X., Liu, B., Liu, W., & Zhou, W. (2023). *Adversarial machine learning: attack surfaces, defence mechanisms, learning theories in artificial intelligence*.

Johnson, E., Seyi-Lande, O. B., Adeleke, G. S., Amajuoyi, C. P., & Simpson, B. D. (2024). Developing scalable data solutions for small and medium enterprises: Challenges and best practices. *World Journal of Advanced Research and Reviews*, 22(3), 1910–1935.

Efe, A., & Geliş, M. (2024). Risk Modelling of Cyber Threats Against MIS and ERP Applications. *Pamukkale University Journal of Business Research*, *11*(2), 502–530.

Al-Wajih, J. A. (2023). Security challenges in cloud computing: A comprehensive analysis. *Journal of Advanced Engineering*, 2023(2), 155–181.

Ferdous, J., Islam, R., Mahboubi, A., Access, M. I.-Iee., & 2023, undefined. (n.d.). A review of state-of-the-art malware attack trends and defense mechanisms. *Ieeexplore.Ieee.OrgJ Ferdous, R Islam, A Mahboubi, MZ IslamIEEe Access, 2023•ieeexplore.Ieee.Org*. Retrieved May 2, 2026.

Gong, X., Wang, Q., Chen, Y., Yang, W., & Jiang, X. (2020). Model Extraction Attacks and Defenses on Cloud-Based Machine Learning Models. *IEEE Communications Magazine*, *58*(12), 83–89.

Iglesias, F., Ferreira, D., Vormayr, G., Bachl, M., Sciences, T. Z.-A., & 2020, undefined. (n.d.). NTARC: a data model for the systematic review of network traffic analysis research. *Mdpi.ComF Iglesias, DC Ferreira, G Vormayr, M Bachl, T ZsebyApplied Sciences, 2020•mdpi.Com*. Retrieved May 3, 2026.

Jada, I., Management, T. M.-D. and I., & 2024, undefined. (n.d.). The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. *Elsevier*. Retrieved May 2, 2026.

Jha, K. M. (2022). Exploring the Role of Neural Networks in Big Data-Driven ERP Systems for Proactive Cybersecurity Management . *SSRN Electronic Journal*.

Kechagias, E. P., Chatzistelios, G., Papadopoulos, G. A., & Apostolou, P. (2022). Digital transformation of the maritime industry: A cybersecurity systemic approach. *International Journal of Critical Infrastructure Protection*, *37*, 100526.

Khalaf, N. Z., Al Barazanchi, I. I., Al Barazanchi, I. I., Radhi, A. D., Radhi, A. D., Shah, P., Sekhar, R., Khalaf, N. Z., Barazanchi, I. I. Al, Barazanchi, I. I. Al, Radhi, A. D., Radhi, A. D., Shah, P., & Sekhar, R. (2025). Development of real-time threat detection systems with AI-driven cybersecurity in critical infrastructure. *Mesopotamian Journal of CyberSecurity*, *5*(2), 501–513.

Kilincer, I., Ertam, F., Networks, A. S.-C., & 2021, undefined. (n.d.). Machine learning methods for cyber security intrusion detection: Datasets and com-

parative study. *ElsevierIF Kilincer, F Ertam, A SengurComputer Networks, 2021•Elsevier*. Retrieved May 3, 2026.

Kohar, A., Rizky Muhammad Hendrik Noor Asegaff, A., Sebastian Salim, B., Wijaya, H., Rakhmad, H., & Kalimantan Muhammad Arsyad Al Banjari, I. (2026). Evaluasi Kinerja Algoritma Deep Learning Untuk Deteksi Dini Serangan Siber Pada Jaringan Komputer. *Jurnal Pengabdian Masyarakat Dan Riset Pendidikan*, *4*(3), 14602–14607.

Kumar Bhardwaj, A., Dutta, P. K., Chintale, P., & History, A. (2024). Securing container images through automated vulnerability detection in shift-left CI/CD pipelines. *Mesopotamian.PressAK Bhardwaj, PK Dutta, P Chintale-Babylonian Journal of Networking, 2024•mesopotamian.Press*, *2024*, 162–170.

Lahare, P. A., & Wakchaure, M. A. (2025). Proactive defense through automated cyber threat detection and intelligence: Latest trends and challenges. *AIP Conference Proceedings*, *3327*(1).

Lehto, M. (2022). Cyber-Attacks Against Critical Infrastructure. *Computational Methods in Applied Sciences*, *56*, 3–42.

Madhav Jha, K., Bodepudi, V., Babu Boppana, S., Katnapally, N., Rao Maka, S., & Sakuru, M. (n.d.). *Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems*. *22*(1), 2023–6193.

Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems 2025 67:8*, *67*(8), 6969–7055.

Muhati, E., Privacy, D. R.-J. of C. and, & 2024, undefined. (n.d.). Data-driven network anomaly detection with cyber attack and defense visualization. *Mdpi.ComE Muhati, D RawatJournal of Cybersecurity and Privacy, 2024•mdpi.Com*. Retrieved May 3, 2026.

Nugroho, S. A., Sumaryanto, S., & Hadi, A. P. (2024). The Enhancing Cybersecurity with AI Algorithms and Big Data Analytics: Challenges and Solutions. *Journal of Technology Informatics and Engineering*, *3*(3), 279–295.

Okoli, U., Obi, O., & Atuegwu, A. (2024). Machine learning in cybersecurity: A review of threat detection and defense mechanisms. World Journal of Advanced Engineering Research and Science, 11(4), 45–58.

Olaoye, G. (2025). *Exploring the Role of Neural Networks in Big Data-Driven ERP Systems for Proactive Cybersecurity Management*.

Omotoye, S., & Chen, W. (2026). *Development of a Comprehensive Framework for Detecting Insider Threats*. 379–389.

Raja, J. A., Khang, A., & Vani, R. (2024). Ransomware resilience strategies for manufacturing systems: Safeguarding the enterprise resource planning and human resource management data. *Machine Vision and Industrial Robotics in Manufacturing: Approaches, Technologies, and Applications*, 435–448.

Reddy, S. P. K., Nagavelli, U., Kiran, Y. S., Kondoju, C. S., Bushmoni, S., & Yashaswi, A. (2024). Deep Learning for Zero-Day Threat Detection and Mitigation. *Proceedings of 5th International Conference on IoT Based Control Networks and Intelligent Systems, ICICNIS 2024*, 1362–1368.

Saeed, S., Altamimi, S. A., Alkayyal, N. A., Alshehri, E., & Alabbad, D. A. (2023). Digital Transformation and Cybersecurity Challenges for Businesses Resilience: Issues and Recommendations. *Sensors 2023, Vol. 23, Page 6666*, *23*(15), 6666.

Shaheen, A. (2023). Cybersecurity in the Modern Era: An Overview of Recent Trends. *Journal of Engineering and Computational Intelligence Review*, *1*(1), 39–50.

Subramanian, N., Engineering, A. J.-C. & E., & 2018, undefined. (n.d.). Recent security challenges in cloud computing. *Elsevier*. Retrieved May 3, 2026.

Younis, R., Iqbal, M., Munir, K., Javed, M. A., Haris, M., & Alahmari, S. (2024). A Comprehensive Analysis of Cloud Service Models: IaaS, PaaS, and SaaS in the Context of Emerging Technologies and Trend. *5th International Conference on Electrical, Communication and Computer Engineering, ICECCE 2024*.

Bölüm 6

İşitsel ve Görsel Verilerle Ruhsal Bozuklukların Hesaplamalı Analizinde Veri İşleme Hatları, Öznitelik Çıkarımı ve Çok-Kipli Füzyon

Uygar Aydin[1]

İnci Zaim Gökbay[2]

Özet

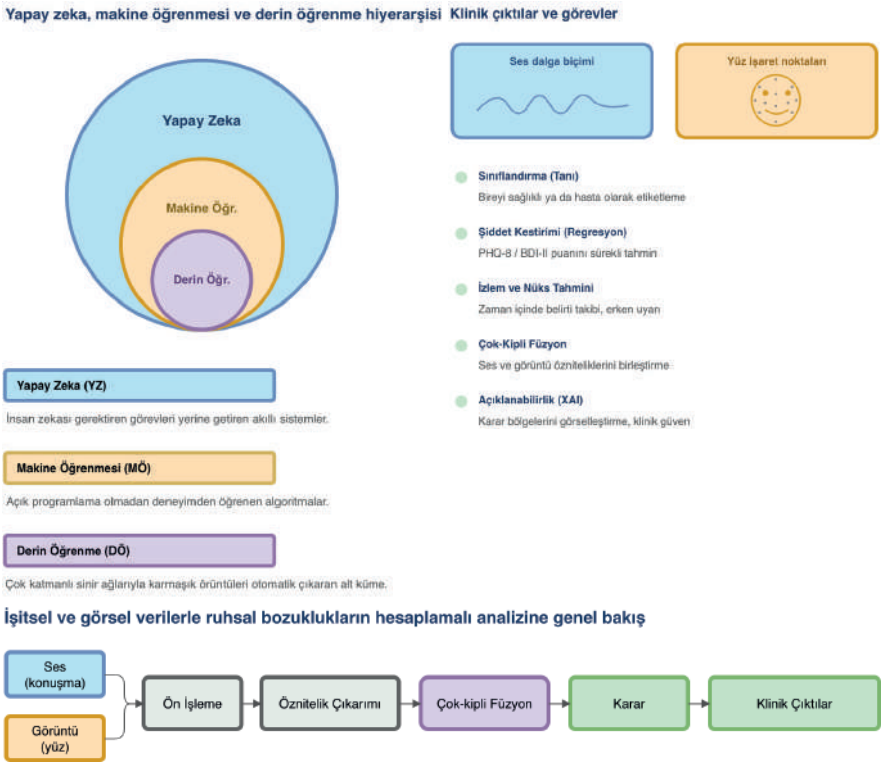
Ruhsal bozuklukların erken tanısı, tedavisi ve izlenmesi, belirtilerin öznel doğası ve geleneksel klinik yöntemlerin ölçüm sınırlılıkları nedeniyle güçtür. Bu bölüm, işitsel, görsel ve çok-kipli verilerden nesnel biyobelirteçler elde etmeyi amaçlayan hesaplamalı yaklaşımları sinyal işleme ve makine öğrenmesi perspektifinden ele almaktadır. Veri kaynakları, klinik değerlendirme ölçekleri, ön işleme adımları, ses ve görüntü verileri için öznitelik çıkarım yöntemleri, sınıflandırma mimarileri ve çok-kipli füzyon stratejileri teknik düzeyde incelenmektedir. Bu teknik çerçeve son yıllarda belirli yöntemler ve hedefler çevresinde yoğunlaşan ve özellikle COVID-19 sonrası dönemde hızla genişleyen kapsamlı bir literatüre dayanmaktadır. Bu yöntem ve hedef yoğunlaşması içinde depresyon tespiti baskın araştırma hedefi olarak öne çıkmış, evrişimli sinir ağları (CNN) ise temel mimari haline gelmiştir. Öznitelik düzeyinde gerçekleştirilen ses-görüntü füzyonu, tek-kipli çözümlere kıyasla kayda değer doğruluk kazanımları sağlamıştır. Bölümde ayrıca temsili yüz ve ses tanıma mimarileri, temel ve yardımcı sınıflandırma yöntemleri arasındaki ayrım ile uzaktan ve sürekli değerlendirme, mahremiyeti koruyan girişimsel olmayan izlem ve kaynak erişiminin sınırlı olduğu koşullara uyarlanabilirlik gibi gerçek yaşam uygulamaları tartışılmaktadır. Bununla birlikte kültürel ve dilsel çeşitlilikten yoksun veri kümeleri, tanı ile tedavi arasındaki boşluk ve modellerin yorumlanabilirlik eksikliği alandaki yeniliklerin klinik etkiye dönüşmesinin önündeki başlıca engeller olmayı sürdürmektedir. Dolayısıyla bu

1 İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Anabilim Dalı, İstanbul; uygaraydin1@ogr.iu.edu.tr; https://orcid.org/0000-0002-9052-3512

2 İstanbul Üniversitesi, Bilgisayar ve Bilişim Teknolojileri Fakültesi, Yapay Zeka ve Veri Mühendisliği Bölümü, Yapay Zeka ve Veri Mühendisliği Anabilim Dalı, İstanbul; inci.gokbay@istanbul.edu.tr; https://orcid.org/0000-0002-4488-1642

https://doi.org/10.58830/ozgur.pub1351.c5539

engellerin aşılması alanın önceliklerini doğrudan şekillendirmektedir. Yapılan araştırmalar model doğruluğunu artırmanın ötesinde, bu alanda çeşitlilik içeren ve uzun vadeli veri kümeleri oluşturulmasının, modellerin tüm klinik sürecini kapsayacak şekilde genişletilmesinin ve klinik ihtiyaçları doğrudan karşılayan, yorumlanabilir sistemler geliştirilmesinin gerekli olduğunu ortaya koymuştur.



Grafik Özet. İşitsel-görsel verilerle ruhsal bozuklukların hesaplamalı analizine genel bakış, yöntem hiyerarşisi, klinik çıktılar ve görevler. (Özgün şekil; yazarlar tarafından oluşturulmuştur.)

1. Giriş

Dünya Sağlık Örgütü (WHO) ruh sağlığını, bireyin yeteneklerinin farkında olduğu, olağan yaşam stresleriyle baş edebildiği, verimli biçimde çalışabildiği ve topluluğuna katkı sunabildiği bir iyilik hali olarak tanımlar (WHO, 2022b). Buna karşın ruhsal sorunlar, fiziksel hastalıklar kadar görünür olmamaları nedeniyle giderek büyüyen bir halk sağlığı yüküne dönüşmüştür. Global Hastalık Yükü (Global Burden of Disease, GBD) 2019 analizine göre dünyada her sekiz kişiden biri (yaklaşık 970 milyon birey) bir ruhsal bozuklukla yaşamaktadır. Bunların yaklaşık 280 milyonu depresyon, 301

milyonu ise anksiyete bozukluğu tanısı taşımaktadır (GBD 2019 Mental Disorders Collaborators, 2022; WHO, 2022c). Ruh sağlığı sorunlarının oluşturduğu yükün ekonomik boyutu da çarpıcıdır. Nitekim İngiltere'de pandemi öncesinde ruhsal bozukluklar sebebiyle ortaya çıkan yıllık ekonomik ve toplumsal maliyetin 105 milyar sterline ulaştığı raporlanmıştır (Adcock ve Parkin, 2016). Bu maliyetin gerisinde, yalnızca bozukluğun kendisi değil, ona erişimi ve tedaviyi güçleştiren etkenler de yer almaktadır. Ruhsal bozukluklar bireyin iyilik hali, eğitim ve çalışma yaşamı ile sosyal ilişkileri üzerinde olumsuz etkiler doğurur. Tanı ve tedavi süreçlerine erişimdeki güçlükler ve uzun bekleme süreleri ise bu etkileri ağırlaştırmakta ve sağlık sistemleri üzerindeki yükü daha da artırmaktadır. Nesnel, ölçeklenebilir ve erişilebilir değerlendirme araçlarına duyulan ihtiyaç bu yapısal sorunlar çerçevesinde belirginleşmektedir.

COVID-19 salgını hem talebi nicel olarak artırmış hem de talebin doğasını değiştirmiştir. Salgının ilk yılında küresel anksiyete yaygınlığında %25,6 ve depresyon yaygınlığında %27,6 oranında bir artış bildirilmiştir (Santomauro vd., 2021; WHO, 2022a). Yüz yüze terapi modellerinin kesintiye uğraması, uzaktan erişilebilir ve ölçeklenebilir dijital çözümlere yönelik acil bir ihtiyaç doğurmuştur. Bu ihtiyaç, duygusal hesaplama (affective computing) araştırmalarının ulaştığı teknolojik olgunlukla birleşince, ses ve yüz temelli hesaplamalı yöntem araştırmalarını belirgin biçimde hızlandırmıştır (He, Niu, vd., 2022; Low vd., 2020). Bu dönemde ulusal ve uluslararası fon kuruluşlarının dijital ruh sağlığı teknolojilerine sağladığı büyük ölçekli kaynaklar, yapay zeka ve makine öğrenmesi uzmanlarının dikkatlerini bu alana çekmiştir. Böylece toplumsal talep, teknolojik olgunluk ve kurumsal teşvik alanın hızla genişlemesinin zeminini hazırlamıştır.

Geleneksel psikiyatrik değerlendirmenin temel kısıtı büyük ölçüde öznel belirti bildirimine ve klinik gözleme dayanmasıdır. Bu noktada konuşma sinyalleri ve yüz ifadeleri gibi gözlemlenebilir davranışsal ve fizyolojik ipuçlarından nesnel biyobelirteçler (biomarker) çıkarmayı amaçlayan hesaplamalı yöntemler önem kazanmıştır (He, Niu, vd., 2022; Low vd., 2020). Bu bölüm, alanı bir mühendislik problemi olarak ele alır ve üç temel eksende yapılandırır. İlk olarak yüz ve ses analizlerinin erken tanıya katkıları, ardından makine ve derin öğrenme tekniklerinin tanı ile tedaviyi desteklemedeki etkinliği, son olarak da bu tekniklerin tanı, tedavi ve takip sürecini iyileştirme yolları değerlendirilmektedir.

Bu bölümün ele aldığı yaklaşımlar salt teknik çözümler değil, daha geniş bir kuramsal dönüşümün parçasıdır. Yapay zeka ve özellikle derin öğrenme, elle tasarlanmış kurallar yerine temsilleri doğrudan veriden hiyerarşik biçimde öğrenmeyi mümkün kılarak görüntü, ses ve dil işlemede paradigma değiştirici

bir rol üstlenmiştir (LeCun vd., 2015). Bu modellerin gücü büyük ölçüde veri ölçeğine bağlı olduğundan yüksek hacim, hız ve çeşitlilik gösteren veri kümelerini niteleyen büyük veri (big data), sağlık ile davranış bilimlerinde giderek merkezi bir rol kazanmıştır (Monteith vd., 2015). Sağlık ve davranış verisinin dijitalleşmesiyle ortaya çıkan bu veri yığınları, öğrenme temelli yöntemlerin itici gücü haline gelmiştir. Bu öğrenme paradigması ve büyük veri birikimi, ruh sağlığı alanında iki gelişmede somutlaşır. Birincisi hesaplamalı psikiyatrinin (computational psychiatry) gelişimidir. Beyin ve davranışı matematiksel modellerle ele alan bu yaklaşım, nörobilim ile klinik uygulama arasında veri ve kuram temelli bir köprü kurar (Montague vd., 2012; Huys vd., 2016). İkincisi ise dijital fenotiplemedir (digital phenotyping). Akıllı telefon ve giyilebilir cihazların yaygınlaşması bireyin günlük davranışını sürekli ve nesnel biçimde ölçen büyük ölçekli veri akışları doğurarak bu kavramı ortaya çıkarmıştır (Insel, 2017). Büyük veri, öğrenme paradigması ve klinik gereksinimin kesişiminde olan makine öğrenmesi, psikiyatrik değerlendirmeyi öznel bildirimden nesnel ölçüme taşıma potansiyeli taşır (Dwyer vd., 2018). İşitsel ve görsel veriler, davranışsal bilgi açısından zengin ve en az girişimsel kaynaklar arasında yer aldığından bu bölümün odağını oluşturur.

Bu odak doğrultusunda bölümde veri kaynakları, veri işleme hattı, performans değerlendirmeleri ve füzyon stratejileri, gerçek yaşam uygulamaları ve klinik entegrasyon, tartışma ve sınırlılıklar ile sonuç ve öneriler derinlemesine incelenmiştir.

2. Veri Kaynakları ve Klinik Ölçekler

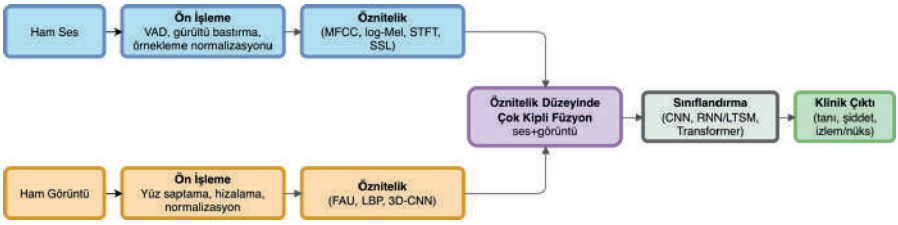
Bu alandaki çalışmalar hem açık kaynak hem de araştırmacıların kendileri tarafından toplanmış olan veri kümelerine dayanır. En sık kullanılan açık kaynak kümeleri klinik görüşme videoları içeren DAIC-WOZ (Gratch vd., 2014) ve AVEC serisidir (Valstar vd., 2013; Valstar vd., 2016). Bu kümeler çok-kipli ve doğrudan ruhsal bozukluklara odaklandıkları için tercih edilirler. DAIC-WOZ veri kümesi bir sanal görüşmecinin yürüttüğü yarı yapılandırılmış klinik görüşmelerden oluşur ve PHQ-8 etiketleriyle birlikte sunulur. AVEC serisi ise 2013'ten bu yana depresyon tanısını bir alt yarışma olarak ele almış ve alana standart bir kıyaslama zemini kazandırmıştır (Valstar vd., 2013; Valstar vd., 2016). Bunların yanında çok-kipli MODMA, etkileşimli duygusal IEMOCAP ve vlog temelli D-Vlog gibi kümeler de kullanılmış, yalnızca görüntü içeren CK+, yalnızca ses içeren ISED ve RAVDESS (Livingstone ve Russo, 2018) ile metin temelli ISEAR gibi daha özelleşmiş kaynaklar tekil çalışmalarda yer bulmuştur. Hastalık şiddeti ağırlıklı olarak Hasta Sağlığı Anketi (Patient Health Questionnaire, PHQ-8) ve Hamilton Depresyon Derecelendirme Ölçeği (Hamilton Depression Rating Scale, HAMD) gibi standart klinik ölçeklerle

etiketlenmiştir. Bazı çalışmalarda mevcut ölçeklerin tanı için yetersiz kaldığı gerekçesiyle Zung Kendini Değerlendirme Depresyon Ölçeği (Self-Rating Depression Scale, SDS) gibi alternatifler tercih edilmiştir (Xie vd., 2021).

Mevcut veri kümelerinin sunduğu zaman tasarrufu ve etik kolaylığa karşılık, bu kümeler özelleştirilemez ve belirli dillerle sınırlıdırlar. Bu nedenle bazı araştırmacılar psikiyatrik değerlendirme sırasında çekilen video kayıtlarından kendi kümelerini oluşturmuştur. Örnekler arasında yüksek çözünürlüklü gerçek zamanlı analiz için toplanan kümeler (Gilanie vd., 2022), duygusal uyaran sunumuna dayalı görevlerle desteklenen tasarımlar (Liu vd., 2024) ve farklı dil ve popülasyonlara yönelik kümeler (Hall vd., 2024; Kim vd., 2023; Mahayossanunt vd., 2023) yer alır. Bu eğilim, aşağıda tartışılacağı üzere, dilsel ve kültürel çeşitlilik açısından önemli bir sınırlılık yaratmaktadır. Kendi verisini toplayan araştırmacılar uyaran temelli görevler ya da yüksek çözünürlüklü kayıt düzenekleriyle depresif belirtileri daha belirgin biçimde ortaya çıkarmayı amaçlamıştır. Bu çabalar değerli olmakla birlikte, örneklem sınırlılığı ve standart dışı protokoller sonuçların karşılaştırılabilirliğini ve yeniden üretilebilirliğini sınırlandırmaktadır.

3. Veri İşleme Hattı

Hesaplamalı yöntemlerin büyük çoğunluğu ön işleme (preprocessing), öznitelik çıkarımı ve sınıflandırmadan oluşan üç aşamalı klasik bir hattı izler. Ses ve görüntü kollarının öznitelik düzeyinde birleştiği bu üç aşamalı işleme hattının genel görünümü Şekil 1'de gösterilmektedir.



Şekil 1. İşitsel-görsel ruhsal bozukluk analizinde veri işleme hattı; ses ve görüntü kolları öznitelik düzeyinde birleşir.

3.1. Ön İşleme

Ön işleme, ham ses ve video kayıtlarını öznitelik çıkarımına uygun, gürültüden arındırılmış ve normalize edilmiş bir temsile dönüştürür. Tipik adımlar konuşma etkinliği tespiti (Voice Activity Detection, VAD), yüz saptama ve hizalama, çerçeve seçimi ile ölçek ve aydınlatma normalizasyonunu içerir. Gerçek-zamanlı uygulamalarda yüksek video çözünürlüğü ince kas hareketlerinin

ve mikro-ifadelerin (micro-expression) daha belirgin yakalanmasına olanak tanır (Gilanie vd., 2022). Ses tarafında bu hazırlık gürültü bastırma, sessiz bölümlerin konuşma etkinliği tespitiyle ayıklanması ve örnekleme oranının standartlaştırılmasını kapsar. Görüntü tarafında ise yüz işaret noktalarına (facial landmarks) göre hizalama, ilgi bölgesinin kırpılması ve poz ile aydınlatma farklarının dengelenmesi öne çıkar. Bu adımlardaki bir hata sonraki katmanlara doğrudan taşındığından, ön işleme kalitesi nihai sınıflandırma başarısının görünmeyen ama belirleyici bir bileşenidir.

3.2. Öznitelik Çıkarımı

Öznitelik çıkarımı, el yapımı öznitelikler ve öğrenilmiş derin temsiller olmak üzere iki temel yaklaşıma dayanır. Görüntü tarafında öğrenilmiş derin temsil yaklaşımı kendi içinde ikiye ayrılır. Bir uçta ResNet, VGG, SeNet ve üç boyutlu evrişimli ağlar gibi mimariler ham görüntüden uçtan uca öznitelik öğrenir. Diğer uçta OpenFace ve FaceReader gibi araç takımları iki aşamalı bir hat kurar. Bu araçlar görüntüden önce yüz işaret noktaları, bakış yönü ve yüz hareket birimi yoğunlukları gibi yorumlanabilir orta düzey öznitelikler çıkarır, sonraki ağ ise bu zenginleştirilmiş temsiller üzerinde çalışır. Yorumlanabilir çıktıları nedeniyle OpenFace ve FaceReader en sık başvurulan araçlar arasında yer alır. Bu ikinci yol alanın ham pikselleri tek aşamada işlemekten çok aşamalı ve yüksek düzeyli temsillere yönelimini yansıtır. Aynı yönelimin bir uzantısı olarak derin mimarilere dikkat mekanizması (attention mechanism) ve çok ölçekli temsil üreten Özellik Piramit Ağları (Feature Pyramid Networks) gibi bileşenler eklenerek daha soyut ve göreve özgü temsiller hedeflenir (Xu vd., 2024). Uçtan uca öğrenme ile araç temelli yüksek düzeyli temsil çıkarımı arasındaki seçim, yorumlanabilirlik (interpretability) ile esneklik arasındaki dengeye ve veri büyüklüğüne göre yapılır. Uçtan uca derin ağlar daha esnek ancak yorumlanması güç temsiller üretirken, araç temelli betimleyiciler daha yorumlanabilir olup az veriyle daha kararlı sonuç verir.

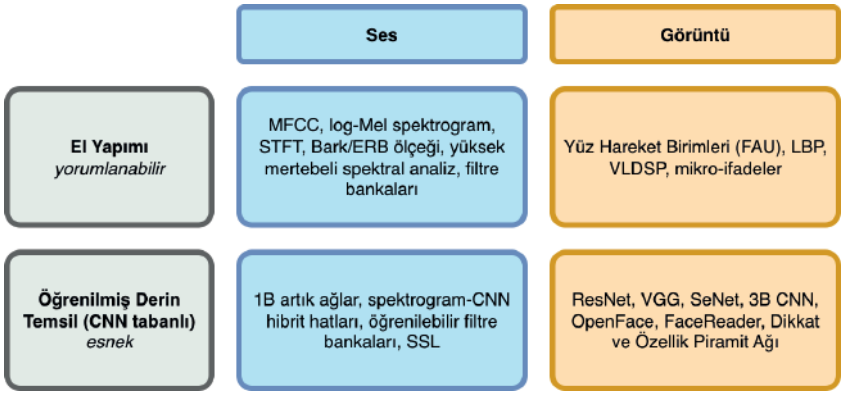
El yapımı görüntü betimleyicileri çoğunlukla bu yorumlanabilir uca konumlanır. Bunların başında, yüz ifadelerinin anatomik temelini doğrudan yansıttığı için yüksek açıklanabilirliğe sahip Yüz Hareket Birimleri (Facial Action Units, FAU) gelir. Nitekim OpenFace gibi araçların ürettiği temel çıktı da bu birimlerdir. Doku temelli Yerel İkili Örüntüler (Local Binary Patterns, LBP) ve dinamik bir betimleyici olan Hacimsel Yerel Yönelimli Yapısal Örüntü (Volume Local Directional Structural Pattern, VLDSP) (Uddin vd., 2022) ile mikro-ifade analizine dayalı yöntemler (Gilanie vd., 2022) bu aileyi tamamlar. Bu çeşitlilik, el yapımı betimleyicilerin yorumlanabilirliği ile öğrenilmiş temsillerin esnekliğini birleştiren hibrit tasarımlara olan ilgiyi yansıtır.

Ses tarafında benzer bir ayrım geçerlidir. Ham dalga biçimini uçtan uca işleyen 1B artık ağların yanı sıra, el yapımı akustik öznitelikleri girdi alan hibrit veri işleme hatları (pipeline) yaygındır ve bu hatlar daha az veriyle daha kararlı sonuç hedefler. El yapımı akustik öznitelikler arasında Mel-Frekansı Kepstral Katsayıları (Mel-Frequency Cepstral Coefficients, MFCC) en yaygın spektral temsildir ve sıklıkla CNN tarafından üretilen spektrogram öznitelikleriyle birleştirilir (Das ve Naskar, 2024). COVAREP gibi araç takımları standart akustik öznitelik kümeleri sağlarken (Degottex vd., 2014), öğrenilebilir zaman-uzamı filtre bankaları (learnable time-domain filterbanks) sabit betimleyiciler yerine veriden optimize edilen bir ara çözüm sunar (Yang vd., 2023). Bu çekirdek temsillerin yanında Log-Mel spektrogram, Bark ölçeği ve eşdeğer dikdörtgensel bant genişliği gibi düşük frekansa odaklı temsiller ile Kısa-Zamanlı Fourier Dönüşümü (Short-Time Fourier Transform), Kepstrum, Gabor dönüşümü ve yüksek mertebeli spektral analiz (Higher-Order Spectral Analysis) gibi dönüşüm temelli öznitelikler de kullanılır.

Tablo 1. Ses ve görüntü için başlıca öznitelik çıkarım aileleri ve temsil eden örnek çalışmalar.

Öznitelik ailesi	Başlıca yöntemler	Örnek çalışma
CNN tabanlı görüntü	ResNet, VGG, SeNet, 3B CNN (uçtan uca); OpenFace, FaceReader (iki aşamalı araç); dikkat ve Özellik Piramit Ağı (mimari bileşen)	(Xu vd., 2024)
CNN tabanlı ses	1B artık ağlar (ham dalga biçimi), spektrogram-CNN hibrit hatları, öğrenilebilir filtre bankaları	(Das ve Naskar, 2024)
El yapımı ses	MFCC, log-Mel spektrogram, STFT, filtre bankaları	(Yang vd., 2023)
El yapımı görüntü	Yüz Hareket Birimleri (FAU), LBP, VLDSP, mikro-ifadeler	(Gilanie vd., 2022)

Öznitelik ailelerinin temsil türü (el yapımı, öğrenilmiş) ve kip (ses, görüntü) ekseninde sınıflandırılması Şekil 2'de özetlenmektedir.



Şekil 2. Öznitelik çıkarımı taksonomisi: temsil türü (el yapımı / öğrenilmiş) × kip (ses / görüntü).

Bu öznitelik ailelerinin somut mimari karşılıkları literatürde belirgindir. Yüz dinamiklerini değerlendirmek için değişken çekirdek boyutları kullanan Çok-Ölçekli Uzamsal-Zamansal Ağ (Multiscale Spatiotemporal Network, MSN), tekdüze çekirdekli C3D gibi modellere kıyasla ince yüz değişimlerini daha verimli yakalamıştır (de Melo vd., 2020). DepNet ise video temelli analizde yüz ifadelerinin zamansal özniteliklerini modelleyerek doğruluğu artırmıştır (He, Guo, vd., 2022). Yorumlanabilirlik açısından kritik bir örnek, küresel ortalama havuzlama (global average pooling) katmanı içeren bir derin evrişimli ağ ile Depresyon Etkinleştirme Haritaları (Depression Activation Maps) üreten DepressNet'tir. Bu haritalar, depresyon şiddetine dair anlamlı bilgi taşıyan yüz bölgelerini işaretleyerek sonuçların klinik yorumunu güçlendirmiştir (Zhou vd., 2020). Ses tarafında sesli ve sessiz harf düzeyinde fonem temelli bir CNN mimarisi olan AudVowelConsNet (Muzammel vd., 2020) ile üç boyutlu evrişimli ağlar (Wang vd., 2021) konuşmanın klinik açıdan ayırt edici örüntülerini yakalamaya yönelik tamamlayıcı yaklaşımlardır.

3.3. Sınıflandırma Mimarileri

Sınıflandırma katmanında evrişimli sinir ağları (Convolutional Neural Network, CNN) baskın mimaridir. CNN, görüntü ve ses verisinde otomatik öznitelik çıkarımı ve sınıflandırma için temel araç haline gelmiştir. Dikkat mekanizmaları (attention mechanism), modelin odağını girdideki ayırt edici bölgelere yönlendirerek daha derin ve ayrıntılı temsiller üretir ve giderek daha sık tamamlayıcı bileşen olarak kullanılır (Othmani vd., 2022; Xu vd., 2024). Donmuş duygulanım veya konuşma hızındaki yavaşlama gibi zamansal dinamiklerin modellenmesinde ise yinelemeli ağlar (Recurrent Neural Network, RNN, LSTM, Bi-LSTM) öne çıkar (Uddin vd., 2022). Genel eğilim, statik

örüntüleri yakalayan temel modellerden, zamansal dinamikleri modelleyen ağlara ve nihayetinde kararını gerekçelendirebilen açıklanabilir yapay zeka (Explainable AI, XAI) yaklaşımlarına doğru bir evrimdir (Mahayossanunt vd., 2023; Xie vd., 2021). Bu mimari çeşitliliğin altında pratik bir gerekçe yatar. Evrişimli ağlar görece az veriyle güçlü uzamsal öznitelikler öğrenebildikleri için baskın omurgayı oluştururken, yinelemeli ağlar ardışık çerçeveler arasındaki bağımlılıkları modelleyerek konuşma temposu ve ifade geçişleri gibi zamansal ipuçlarını yakalar. Evrişimli ve yinelemeli ağların bu rol paylaşımı dışında kalan mimariler de belirli sınırlılıkları aşmak üzere denenmiştir. Çizge sinir ağları (Graph Neural Networks, GNN) ve karışım modelleri kipler arası ilişkileri daha esnek temsil edebildikleri için niş senaryolarda sınanmış, Transformer (dönüştürücü) mimariler ise yinelemeli ağların zorlandığı uzun menzilli bağımlılıkları dikkat ağırlıklarıyla modelleyerek yeni bir yön açmıştır. Her farklı mimari belirli bir ihtiyaca yanıt verdiğinden mimari seçimi tek bir ölçüte değil; veri büyüklüğü, yorumlanabilirlik beklentisi ve hesaplama bütçesi arasındaki dengeye dayanır.

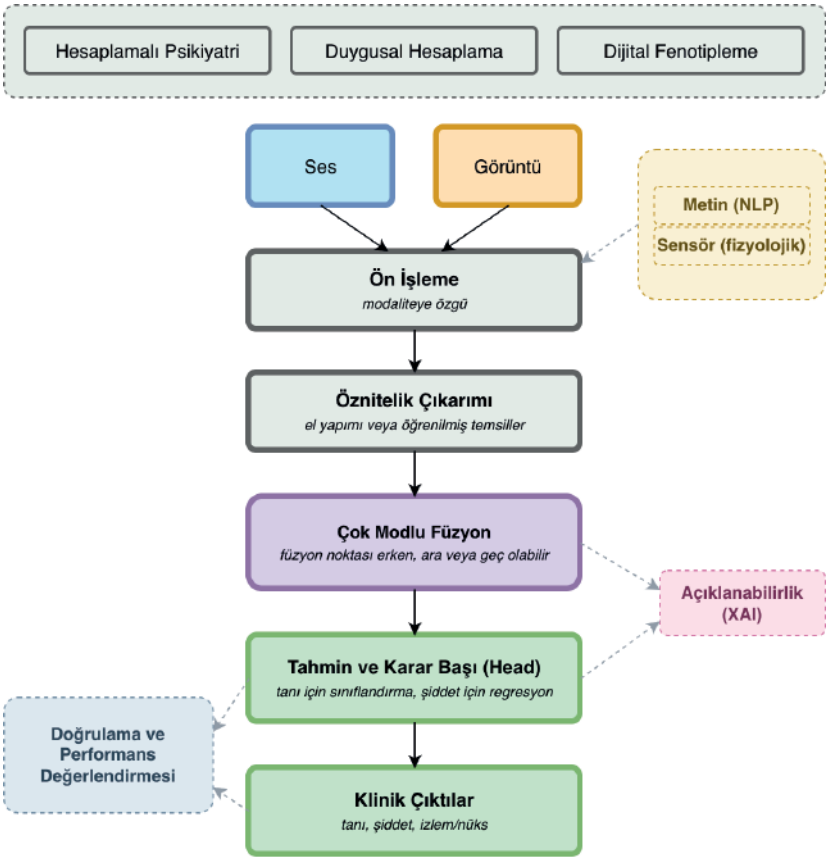
Bu mimariler tek tek ele alınmalarının yanı sıra, tanı sürecinde üstlendikleri role göre de sınıflandırılabilir. Tanı desteği sağlayan bu modeller yardımcı bir mekanizma kullanıp kullanmamalarına göre iki gruba ayrılabilir. Temel yaklaşımlar, tek başına bir CNN veya geleneksel makine öğrenmesi modeliyle doğrudan sınıflandırma yapar. Yardımcı yaklaşımlar ise bu omurgayı dikkat mekanizmaları, uzamsal-zamansal modüller veya çok-kipli füzyon (multimodal fusion) ile güçlendirir. Örneğin yüz ifadesi ile göz bebeği tepkisini bir öz-dikkat (self-attention) ağında birleştiren (Liu vd., 2024) ya da dikkat mekanizmasını bir öznitelik piramidiyle eşleştiren (Xu vd., 2024) çalışmalar, ayırt edici bölgelere odaklanarak başarımı artırmayı amaçlar. Söz konusu ayrım, model karmaşıklığı ile yorumlanabilirlik ve veri verimliliği arasındaki ödünleşimi de görünür kılar. Pratikte yardımcı mekanizmaların katkısı, eklenen karmaşıklığın getirdiği veri ve hesaplama maliyetiyle birlikte ve veri büyüklüğüne göre değerlendirilmelidir. Nitekim küçük örneklemlerde sade bir omurga, aşırı uyum riski nedeniyle daha güvenilir sonuç verebilir.

Ses temelli tanıma yaklaşımları da benzer bir çeşitlilik gösterir. Mel-Frekansı Kepstral Katsayıları ile CNN tarafından üretilen spektrogram özniteliklerini birleştiren melez modeller yaygın bir başlangıç noktasıdır (Das ve Naskar, 2024). Öğrenilebilir zaman-uzamı filtre bankaları ise sabit el yapımı filtreler yerine veriden öğrenilen, dikkatle yönlendirilen temsiller sunar (Yang vd., 2023). Fonem düzeyinde uzmanlaşmış mimariler (sesli ve sessiz harfleri ayrı ağlarla işleyen AudVowelConsNet gibi) konuşmanın ince akustik yapısını hedeflerken (Muzammel vd., 2020), üç boyutlu evrişimli ağlar zaman ve frekans eksenlerini birlikte modelleyerek depresif konuşma örüntülerini

yakalamaya çalışır (Wang vd., 2021). Bu yaklaşımların ortak amacı, depresif konuşmanın düşük perde, monoton tonlama ve yavaşlamış tempo gibi ayırt edici örüntülerini, dile ve konuşmacıya olabildiğince bağımsız bir biçimde temsil etmektir.

Bu CNN-merkezli manzara, son yıllarda öz-denetimli öğrenme (self-supervised learning, SSL) temelli konuşma temel modellerinin (foundation models) yükselişiyle değişmektedir. wav2vec 2.0 (Baevski vd., 2020), HuBERT (Hsu vd., 2021) ve WavLM (Chen vd., 2022) gibi, etiketsiz büyük konuşma veri kümeleri üzerinde önceden eğitilen modeller, sınırlı klinik veriyle dahi güçlü akustik temsiller sağladıkları için el yapımı özniteliklerin ve sıfırdan eğitilen CNN'lerin yerini hızla almaktadır. Bu modeller tipik olarak ince ayar (fine-tuning) veya dondurulmuş gömme (frozen embedding) çıkarımı yoluyla kullanılır ve veri kısıtı olan ortamlarda transfer öğrenme yoluyla başarımı artırdıkları gösterilmiştir (Wu vd., 2023; Zhang vd., 2024). Güncel yönelim, sabit spektral betimleyiciler yerine bu öğrenilmiş temsilleri CNN-Transformer melez mimarileriyle veya dikkat-havuzlamalı yinelemeli katmanlarla birleştirmektir. Dolayısıyla alandaki mimari özet, CNN'in baskın omurga olduğu bir enstantanenin ötesine geçerek, önceden eğitilmiş temel modellerin ve dikkat tabanlı mimarilerin giderek merkezi rol üstlendiği bir geçiş dönemini yansıtmaktadır.

Veri kaynaklarından ön işleme, öznitelik çıkarımı ve sınıflandırmaya uzanan bu işleme zinciri, paradigma katmanı ve açıklanabilirlik bileşenleriyle birlikte Şekil 3'te bütünleşik bir hesaplamalı psikiyatri çerçevesi olarak sunulmaktadır.



Şekil 3. İşitsel ve görsel verilerle ruhsal bozukluk analizinde bütünleşik hesaplamalı psikiyatri çerçevesi

4. Performans Değerlendirmesi ve Çok-Kipli Füzyon

Bir tanı sisteminin başarımı tek bir sayıya indirgenemeyecek kadar çok boyutlu olduğundan, değerlendirme aşamasında birbirini tamamlayan çeşitli ölçütlere başvurulur. Modeller, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi karışıklık matrisi (confusion matrix) temelli ölçütlerin yanı sıra ortalama mutlak hata (Mean Absolute Error, MAE) ve kök ortalama kare hata (Root Mean Square Error, RMSE) ile değerlendirilir. Bunun yanında ROC eğrisi (ROC curve) ve AUC, k-katlamalı çapraz doğrulama (k-fold cross-validation) ile korelasyon katsayıları kullanılır. Klinik bağlamda özgüllük (specificity) özellikle kritiktir. Sağlıklı bir bireyin hasta olarak sınıflandırılması (yanlış pozitif), gereksiz ilaç kullanımına, psikolojik strese ve toplumsal damgalanmaya yol açabilir. Bu nedenle tanısal modellerin yalnızca duyarlılığı değil, yanlış pozitifleri azaltan ölçütleri de gözetmesi gerekir.

Çalışmaların çoğu, doğruluğun yanında özgüllüğü, ROC eğrisi altındaki alanı ve k-katlamalı çapraz doğrulama sonuçlarını birlikte raporlar. Sürekli çıktı üreten modellerde ise Uyum Korelasyon Katsayısı (Concordance Correlation Coefficient, CCC) ve Pearson korelasyonu tercih edilir. Metrik seçimindeki bu bilinçli çeşitlilik, modellerin yalnızca ortalama başarısını değil, farklı alt gruplardaki kararlılığını, genellenebilirliğini ve klinik güvenliğini de görünür kılmayı amaçlar.

Tablo 2. Bölümde atıflı seçilmiş çalışmaların raporlanan başarımı.

Çalışma	Veri kümesi (N)	Kip / Öznitelik	Hedef	Doğrulama	Raporlanan başarım
Othmani vd. (2022)	DAIC-WOZ (189)	Ses + görüntü füzyon (FAU + spektrogram)	İkili (nüks/ depresyon)	LOSO	Doğruluk %87,4; F1 %82,3
Muzammel vd. (2020)	DAIC-WOZ (189)	Ses (fonem-düzeyi CNN)	PHQ-8 ikili	Eğitim-test bölmesi	Doğruluk %86,06; AUC 0,83; ort. F1 %85,85
Das ve Naskar (2024)	DAIC-WOZ; MODMA	Ses (MFCC + CNN-spektrogram)	İkili	Çapraz doğrulama	Doğruluk >%90 (DAIC ve MODMA)
Kim vd. (2023)	Korece, kendi (318)	Ses (log-Mel CNN)	İkili (MDB/ kontrol)	10-katlı CV	Doğruluk %78,14; ort. AUC 0,86
Zhou vd. (2020)	AVEC2013/2014	Görüntü (çok-bölgeli ResNet)	BDI-II regresyon	AVEC protokolü	MAE 6,21; RMSE 8,39 (AVEC2014) †

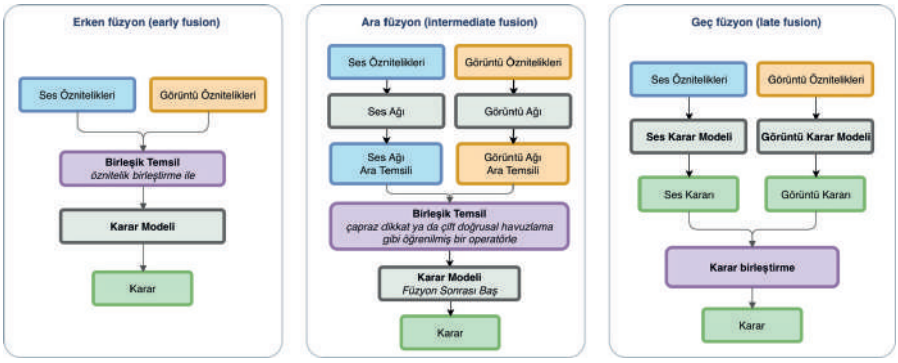
† Zhou vd. (2020) değerleri ikincil karşılaştırma kaynaklarından alınmıştır. Tablodaki çalışmalar farklı hedef, doğrulama stratejisi ve örneklem kullandığından doğrudan bir doğruluk sıralaması olarak okunmamalıdır.

Söz konusu ölçütler tek bir kipe dayalı modellerin ulaştığı sınırları görünür kıldıkça, araştırmacıları farklı veri kiplerini birleştiren füzyon temelli yaklaşımlara yöneltmiştir. Othmani ve arkadaşları, vokal ve görsel ipuçlarına dayalı yeni bir biyobelirteç tanımlayan ve majör depresif bozuklukta nüks olasılığını öngören bir Normallik Modeli (Model of Normality, MoN) çerçevesi geliştirmiştir. Videodan yüz hareket birimi öznitelikleri, konuşmadan ise spektrogram öznitelikleri çıkarılmış ve öznitelik düzeyinde gerçekleştirilen ses-görüntü füzyonu yalnızca sese dayalı modele kıyasla belirgin bir başarım artışı sağlamıştır. En iyi sonuç bir denek dışarıda bırakma (Leave-One-Subject-Out, LOSO) stratejisiyle %87,4 doğruluk ve %82,3 F1 skoru olarak raporlanmıştır (Othmani vd., 2022). Benzer biçimde Uddin ve arkadaşları,

biri ses (LSTM tabanlı) diğeri video için iki uzamsal-zamansal ağ tasarlamış, video ağında özel VLDSP betimleyicisini kullanmış ve öznitelikleri Zamansal Dikkatli Havuzlama ile özetleyip çok-kipli çarpanlara ayrılmış çift doğrusal havuzlama tekniğiyle birleştirmiştir (Uddin vd., 2022). Bu örnekler basit bir öznitelik birleştirmenin dahi çok-kipli modellerin doğruluğunu belirgin biçimde artırabildiğini ortaya koymaktadır. Tanıdan tedavi ve takibe uzanan örnekler sınırlı olmakla birlikte yön göstericidir. Psikotik bozukluklar üzerine yürütülen bir çalışmada yüz ifadeleri FaceReader tabanlı bir duygu tanıma algoritmasıyla çıkarılmış ve ifadelerin zaman içindeki geçişleri Grup Yinelemeli Çoklu Model Tahmini (Group Iterative Multiple Model Estimation, GIMME) ağ modelleriyle incelenmiştir. Psikotik grupta nötr ifadeden mutluluğa geçişlerin belirgin biçimde daha zayıf olduğu Cohen'in d değeriyle raporlanmıştır (Hall vd., 2024). Vokal ve görsel ipuçlarına dayalı Normallik Modeli çerçevesi ise nüks olasılığını öngörerek tanının ötesinde bir izlem perspektifi sunmuştur (Othmani vd., 2022). Bu çalışmalar, modellerin pasif tanı araçlarından dinamik klinik karar destek sistemlerine evrilme potansiyelini somutlaştırmaktadır.

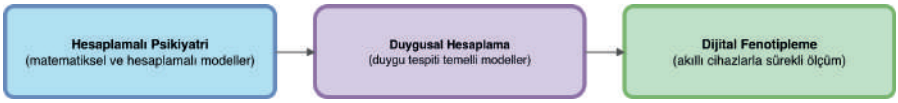
Çok-kipli modellerin bu kazanımlarının ardında kiplerin hangi aşamada birleştirildiğine ilişkin bir tasarım tercihi yatar. Çok-kipli makine öğrenmesinin klasik taksonomisi füzyonu, erken ve geç füzyon ile bu ikisini birleştiren melez (hybrid) füzyon olarak ayırır (Baltrušaitis vd., 2019). Derin öğrenme literatürü ise buna, kiplerin öğrenilmiş ara temsillerini ağın orta katmanlarında birleştiren ara (intermediate) füzyonu ekleyerek füzyonu erken, ara ve geç olmak üzere üç düzeyde ele alır (Stahlschmidt vd., 2022). Erken füzyon (early fusion) ham veriyi ya da düşük düzeyli öznitelikleri tek bir ortak temsilde birleştirip tek bir model eğitir. Kipler arası ince etkileşimleri yakalayabilir ancak zamansal hizalama ve boyut uyumsuzluğuna duyarlıdır. Geç füzyon (late fusion) her kip için ayrı modeller eğitip yalnızca karar düzeyinde, örneğin olasılıkların birleştirilmesiyle bütünleştirir. Kip kaybına karşı dayanıklı ve uygulaması kolaydır, fakat kipler arası tamamlayıcı bilgiyi büyük ölçüde göz ardı eder. Ara füzyon (intermediate fusion) ise kiplerin öğrenilmiş ara temsillerini ağın orta katmanlarında, çoğunlukla dikkat ya da çift doğrusal havuzlama gibi mekanizmalarla birleştirir ve iki uç arasında denge kurar. Othmani ve arkadaşlarının öznitelik düzeyinde birleştirme yaklaşımı erken füzyona, yüz ve göz bebeği temsillerini bir öz-dikkat ağında bütünleştiren tasarım (Liu vd., 2024) ile çok-kipli çift doğrusal havuzlama (Uddin vd., 2022) ise ara füzyona karşılık gelir. Geç füzyona örnek olarak, ses, görüntü ve metin kiplerinin her biri için ayrı eğitilen modellerin tahminlerini ağırlıklı ortalamayla birleştiren AVEC 2016 topluluk sistemi verilebilir (Williamson vd., 2016). Genel eğilim erken ya da geç füzyondan kipler arası etkileşimi öğrenen dikkat tabanlı ara

füzyona doğrudur. Erken, ara ve geç füzyonun şematik karşılaştırması Şekil 4'te verilmektedir.



Şekil 4. Çok-kipli füzyon stratejileri. Erken (öznitelik düzeyi), ara (öğrenilmiş ara temsil düzeyi) ve geç (karar düzeyi) füzyon.

Füzyon stratejilerindeki yönelim COVID-19 salgını sonrasında belirgin biçimde hızlanan ve alanın genelini saran büyüme ve olgunlaşmanın da bir parçasıdır. Bu büyüme alanın ulaştığı teknolojik olgunlukla da uyumludur (He, Niu, vd., 2022; Low vd., 2020). Bu olgunlaşmanın ardında ruh sağlığı bilimlerindeki üç aşamalı yapısal dönüşüm yatar. Bu süreç geleneksel ve belirti merkezli yaklaşımlardan matematiksel ve hesaplamalı modellerle çalışan Hesaplamalı Psikiyatri'ye (computational psychiatry) geçişle başlamıştır. Akabinde duygu tespitine dayalı Duygusal Hesaplama modellerinin yükselişiyle sürmüş ve son olarak akıllı cihazlarla sürekli veri toplamayı mümkün kılan Dijital Fenotipleme (digital phenotyping) çağına ulaşmıştır. Benzer eğilim AVEC yarışma serisinin on yıl içinde temel duygu tanımadan klinik açıdan daha anlamlı problemlere (depresyon ve bipolar bozukluk tanıma) evrilmesinde de görülür (Valstar vd., 2013; Valstar vd., 2016). Şekil 5'te görüldüğü gibi bu üç evre birbiriyle ilişkilidir. Günümüz çalışmaları çoğu zaman her üç paradigmanın araçlarını birlikte kullanır. Nitekim klinik görüşme verisinden öğrenen modeller ile akıllı cihaz akışlarını birleştiren hibrit tasarımlar hem laboratuvar denetimini hem de ekolojik geçerliliği aynı çatı altında toplama eğilimindedir.



Şekil 5. Hesaplamalı ruh sağlığında üç-aşamalı paradigma evrimi.

5. Gerçek-Yaşam Uygulamaları ve Klinik Entegrasyon

Hesaplamalı yöntemlerin önemi klinik akışa entegrasyonlarıyla ölçülür. Literatür üç pratik kullanım alanına işaret etmektedir.

5.1. Uzaktan ve Sürekli Değerlendirme

Yüz yüze görüşmenin pandemi veya coğrafi mesafe nedeniyle mümkün olmadığı durumlarda akıllı telefon ve web tabanlı uygulamalar hasta ile klinisyenin eşzamanlı veya eşzamansız etkileşimine olanak tanır (Uddin vd., 2022; Xie vd., 2021). Yüz ve ses teknolojileriyle sürekli izlem, bireyin ruhsal durumundaki ani değişimleri (intihar riski veya ani duygulanım dalgalanmaları gibi) erken saptayarak hızlı müdahaleye olanak verir (Othmani vd., 2022; Wang vd., 2021). Telefon tabanlı uygulamalar ise günlük ruh hali ve belirti takibini mümkün kılarak hastanın kendi tedavisine etkin katılımını sağlar (Prabhu vd., 2022). Böylece hastayı sürecin edilgen bir nesnesi olmaktan çıkarıp aktif bir paydaşa dönüştürerek tedaviye bağlılığı güçlendirir. Bu uygulamalarda psikiyatrik değerlendirme klinik dışına taşınır ve kişiselleştirilmiş, proaktif bir izlem paradigması doğar. Sürekli izlem klinik ziyaretler arasındaki uzun boşluklarda belirti dalgalanmalarının gözden kaçmasını engelleyerek erken uyarı üretebilir ve tedavi planının zamanında uyarlanmasına olanak tanır.

5.2. Girişimsel Olmayan İzlem ve Mahremiyet

Kamera ve mikrofon temelli yöntemler kan testi veya beyin görüntüleme gibi girişimsel yöntemlere kıyasla daha hızlı, düşük maliyetli ve mahremiyeti koruyan bir alternatif sunar (Gilanie vd., 2022; Prabhu vd., 2022). Bu yaklaşım, damgalanma kaygısı yaşayan hastalar için tedaviye erişimi kolaylaştırır ve geleneksel yöntemlerin yarattığı psikolojik engelleri azaltır. Ses ve video temelli sistemlerin klinik doğruluğu korumanın yanında mahremiyeti önceliklendirmesi, uzun vadeli izlem için güven, erişilebilirlik ve etik temelli yeni bir standardın habercisi olarak değerlendirilebilir (Gilanie vd., 2022). Girişimsel olmayan bu yaklaşım, özellikle düşük sosyoekonomik düzeydeki bölgelerde yüksek maliyetli laboratuvar altyapısına olan bağımlılığı azaltır. Bununla birlikte kamera ve mikrofon verisinin sürekli toplanması, rıza yönetimi ve veri güvenliği konularında dikkatli protokoller gerektirir. Aksi halde mahremiyet avantajı hızla bir risk kaynağına dönüşebilir.

5.3. Farklı Kültürler ve Kısıtlı Kaynaklar

Yüz ifadeleri dil bariyerinden büyük ölçüde bağımsız bir gösterge sunarken (Mahayossanunt vd., 2023), ses temelli analizle birleştirildiğinde farklı kültürel bağlamlarda daha etkili sonuçlar elde edilebilir (Das ve Naskar, 2024). Akıllı

telefon, düşük çözünürlüklü kamera ve temel mikrofon gibi yaygın cihazlarla çalışabilen modeller klinik uzman kaynağının kısıtlı olduğu bölgelerde ruh sağlığı hizmetlerine erişimi artırma potansiyeli taşır (Gilanie vd., 2022; Xie vd., 2021). Bu yönüyle yapay zeka destekli sistemler yalnızca gelişmiş sağlık altyapısına sahip bölgelere değil, kaynakları kısıtlı topluluklara da uyarlanabilir. Düşük bant genişliğine ve sınırlı donanıma uyum sağlayan hafif modeller bu erişilebilirliğin teknik önkoşuludur. Ayrıca yerel dillere ve kültürel ifade biçimlerine uyarlanmış modeller küresel ölçekte adil ve temsil edici bir hizmet sunabilmek için kritik öneme sahiptir.

5.4. Açıklanabilirlik ve Klinik Güven

Klinik kabul için modelin yalnızca doğru değil, gerekçesinin de denetlenebilir olması gerekir. Bu gereksinim, açıklanabilir yapay zeka (XAI) yöntemlerini klinik uygulamaların merkezine taşır (Tjoa ve Guan, 2021). XAI yöntemleri iki ana gruba ayrılır. Birinci grup, kararı sonradan açıklayan post-hoc yöntemlerdir. Bunlar arasında evrişimli ağların hangi görüntü bölgelerine baktığını ısı haritalarıyla gösteren sınıf etkinleştirme haritaları ve Grad-CAM (Selvaraju vd., 2017) ile her bir özniteliğin karara katkısını oyun teorisine dayanan Shapley değerleriyle niceleyen SHAP (Lundberg ve Lee, 2017) öne çıkar. Nitekim bu bölümde ele alınan DepressNet'in ürettiği Depresyon Etkinleştirme Haritaları bu sınıf etkinleştirme soyağacının ruh sağlığına uyarlanmış bir örneğidir (Zhou vd., 2020). İkinci grup, açıklamayı modelin yapısına gömen içsel (intrinsic) yöntemlerdir. Dikkat ağırlıklarının görselleştirilmesi ve bütünleşik gradyanlarla beslenen yorumlanabilir mimariler bu gruba girer ve depresyon tespitinde doğrudan uygulanmıştır (Mahayossanunt vd., 2023; Xie vd., 2021). Klinik bağlamda XAI'nin değeri güven, güvenlik ve düzenleyici olmak üzere üç düzeyde açıklanabilir. Güven düzeyinde, klinisyenin modelin kararını sorgulayıp doğrulamasına olanak tanır. Güvenlik düzeyinde, modelin klinik dışı yapay ipuçlarına dayanıp dayanmadığını ortaya çıkararak yanlılığı görünür kılar. Düzenleyici düzeyde ise biyometrik veriyle çalışan tanı sistemlerinin onay süreçleri için gerekli şeffaflığı sağlar. Bununla birlikte post-hoc açıklamalar, kararın gerçek nedenini her zaman sadık biçimde yansıtmayabileceği, dolayısıyla açıklama yöntemlerinin kendisinin de doğrulanması gerektiği unutulmamalıdır (Tjoa ve Guan, 2021).

6. Tartışma ve Sınırlılıklar

Literatürdeki en belirgin örüntü tematik yoğunlaşmadır. Araştırmalar ağırlıklı olarak depresyona odaklanmış; psikoz çok az çalışılmış (Hall vd., 2024), anksiyete bozuklukları ise neredeyse hiç ele alınmamıştır. Bunun nedenleri arasında depresyonun ses ve yüzde görece kolay yakalanabilen

izler bırakması (düşük vokal ton, monoton prozodi, yavaş konuşma hızı, azalmış mimik ve göz teması) ile AVEC (Valstar vd., 2013; Valstar vd., 2016) ve DAIC-WOZ (Gratch vd., 2014) gibi standart, erişilebilir kümelerin araştırmayı bu yöne çekmesi yer alır. Nitekim DAIC-WOZ anksiyete için de tasarlanmış olsa da pratikte yalnızca depresyon ve TSSB (travma sonrası stres bozukluğu) etiketleri yaygın biçimde kullanılmıştır. Bu yoğunlaşma, depresyon modellerini giderek olgunlaştırırken, anksiyete, bipolar bozukluk ve şizofreni gibi belirtileri kısmen örtüşen durumlar için hesaplamalı yöntemlerin gelişimini zorlaştırmaktadır. Oysa bu bozukluklarda da depresyondakine benzer davranışsal ve fizyolojik göstergeler bulunduğundan tanılar arası ortak örüntüleri yakalayabilen modellere ihtiyaç vardır. Depresyonla klinik açıdan en sık iç içe geçen bozukluklardan biri olan anksiyete, bu örüntülerin modellenmesi bakımından kendine özgü sorunlar barındırır.

Anksiyetenin otomatik tespitini güçleştiren bu sorunlar dört başlık altında toplanabilir:

- Anksiyete belirtileri bağlama bağlı ve epizodiktir. Örneğin, sosyal anksiyetesi olan bireyler düşük stresli ortamlarda normal görünüp yalnızca belirli tetikleyiciler altında klinik belirti gösterebilir. Bu da tek seanslık verinin tanısal geçerliliğini sınırlar.
- Aşırı endişe ve ruminasyon gibi çekirdek belirtiler büyük ölçüde içsel bilişsel süreçlerdir. Depresyonun daha belirgin somatik belirtilerinin aksine, bu gizli nitelikler otomatik tespiti zorlaştırır.
- Eş tanılı (komorbid) depresyon-anksiyete tablolarında, anksiyetenin perdeyi yükseltme eğilimi depresyonun perdeyi düşürme etkisini maskeleyebilir. Bu da algoritmaları yanıltan, karşılıklı örtüşen akustik sinyaller üretir.
- Kamera ve kayıt cihazları bir gözlemci paradoksu (observer paradox) yaratır. Kaydedilmek, sosyal anksiyetenin çekirdeğindeki 'değerlendirilme korkusunu' tetikleyerek ağır belirtili hastaların katılımını engelleyebilir ve veriyi yanlı hale getirebilir.

İkinci önemli sorun tanı-tedavi boşluğudur. Mevcut modeller ağırlıklı olarak tanıya odaklanırken, gerçek klinik ve ekonomik yük, uzun vadeli tedavi yönetimi, belirti izlemi ve nüks riski tahmininde yatar. Bu boşluğun kapatılması üç düzeyde engelin aşılmasını gerektirir. Teknolojik düzeyde, derin öğrenmenin 'kara kutu' niteliği klinisyenlerin kararların ardındaki sinyalleri anlamasını engelleyerek modellerin klinik kabulünü güçleştirir. Bu nedenle yorumlanabilir ve açıklanabilir mimariler kritik önemdedir (Mahayossanunt vd., 2023; Xie vd., 2021). Etik ve yasal düzeyde, yüz görüntüsü ve ses kaydı

gibi biyometrik veriler GDPR ve HIPAA kapsamında oldukça hassastır. Verileri toplama, saklama ve kurumlar arası paylaşım üzerindeki sınırlamalar ham veri paylaşımını kısıtlayarak metodolojik durağanlık riski yaratır. Yöntemsel düzeyde ise kesitsel tasarımlar tedavi yanıtı, belirti seyri ve nüks riski gibi boylamsal soruları yanıtlayamaz. Bunun yerine hastaları haftalar veya aylar boyunca izleyen zaman serisi verilerine ihtiyaç vardır. Anksiyete gibi az çalışılan alanlarda veri sınırlılığını aşmak için bağlama duyarlı protokoller, sanal gerçeklik temelli uyarım ve akıllı cihazlarla pasif algılama önerilmektedir. Depresyon-anksiyete eş tanısı için ise çoklu etiketli ve standartlaştırılmış modelleme çerçeveleri gereklidir.

Üçüncü sorun veri kümelerinde çeşitliliğin eksikliğidir. Yaygın kümeler büyük ölçüde İngilizce ve ağırlıklı olarak batılı popülasyonlardan oluşur. Oysa dil, etnik köken ve kültürün depresyon belirtilerinin görünümünü değiştirebildiği bilinmektedir. Bu boşluğa yanıt olarak kimi araştırmacılar Korece (Kim vd., 2023), Tayca (Mahayossanunt vd., 2023) ve Çince (Yang vd., 2023) veri kümeleri oluşturmuştur. Psikotik bozukluklar üzerine çalışan ekipler ise kendi popülasyonlarına özgü kümeler toplamak zorunda kalmıştır (Hall vd., 2024). Bu durum yetersiz temsil edilen bozukluklar için erişilebilir verinin ne denli sınırlı olduğunu açıkça göstermektedir. Bu açığı kapatmaya yönelik bir başka yaklaşım RAVDESS (Livingstone ve Russo, 2018), CREMA-D (Cao vd., 2014) ve eNTERFACE'05 (Martin vd., 2006) gibi profesyonel oyunculardan oluşturulan duygusal ifade kümelerini örnek alarak mahremiyet ve ham veri erişimi sorunlarını azaltan denetimli kümeler üretmektir. Dil, kültür ve etnik köken açısından dengeli, boylamsal ve çok bozukluklu kümelerin geliştirilmesi modellerin genellenebilirliği için belirleyici olabilir.

Bu sınırlılıkların önemli bir kısmı, alanın büyük veri ve temel model (Foundation Model) çağına geçişiyle yeniden çerçevelenmektedir. Temel modeller etiketsiz devasa veri kümeleri üzerinde öz-denetimli biçimde önceden eğitilen ve çok sayıda alt göreve uyarlanabilen genel amaçlı modellerdir (Bommasani vd., 2021). Tıpta bu yaklaşım, farklı kipleri tek bir modelde birleştiren genel tıbbi yapay zeka vizyonuna doğru ilerlemektedir (Moor vd., 2023). Ruh sağlığı için bu geçiş iki yönlü bir fırsat sunar. Bir yandan konuşma ve görüntü temel modelleri sınırlı klinik veriyle dahi güçlü temsiller sağlayarak veri kısıtı sorununu hafifletebilir ve dijital fenotipleme akışlarındaki büyük ölçekli, etiketsiz veriyi değerlendirilebilir kılabilir (Insel, 2017; Dwyer vd., 2018). Öte yandan bu modeller yeni riskleri de beraberinde getirir. Eğitim verisindeki demografik dengesizlikleri ölçekleyerek pekiştirilebilir, hesaplama maliyetleri nedeniyle kaynakları kısıtlı ortamlar için erişilemez hale gelebilir ve kararlarının yorumlanması güçleştiğinden açıklanabilirlik gereksinimi karşılanamayabilir. Dolayısıyla temel modeller çeşitlilik, mahremiyet ve

yorumlanabilirlik sorunlarını ortadan kaldırmaz. Aksine, bu sorunları daha büyük ölçekte yeniden üretme riski taşıdıklarından dikkatli bir değerlendirme gerektirir.

7. Sonuç ve Öneriler

Bu alandaki, özellikle depresyon tespitindeki hesaplamalı yöntemler kayda değer biçimde olgunlaşmıştır. Ancak bu birikimin klinik etkiye dönüşmesi için aşılması gereken yapısal güçlükler vardır. Bu güçlüklerin aşılmasında dört stratejik yönelim öne çıkmaktadır.

- Hedeflenmiş ve temsil gücü yüksek veri kümeleri geliştirilmelidir. Yalnızca daha çeşitli değil, anksiyete, psikoz ve bipolar bozukluk gibi yeterince temsil edilmeyen durumları hedefleyen, kültürler arası ve boylamsal kümelere öncelik verilmelidir. Federe öğrenme gibi mahremiyet koruyan yaklaşımlar ya da RAVDESS (Livingstone ve Russo, 2018), CREMA-D (Cao vd., 2014) ve eNTERFACE'05 (Martin vd., 2006) gibi kümeleri örnek alarak profesyonel oyuncularla üretilen kümeler etik kısıtlarla bilimsel ilerleme arasındaki gerilimi azaltabilir.
- Araştırma kapsamı klinik sürekliliğe genişletilmelidir. Tek seferlik tanı modellerinden tedavi yanıtını öngören, belirti şiddetini dinamik olarak izleyen ve nüks riskini değerlendiren modellere geçilmelidir (Othmani vd., 2022). Böylece bu teknolojiler kişiselleştirilmiş tedaviyi destekleyen aktif klinik karar destek sistemlerine dönüşebilir.
- Klinik kullanım için yorumlanabilirlik önceliklendirilmelidir. Bir modelin kliniğe geçişi yalnızca doğruluğuna değil, güvenilirliğine ve şeffaflığına da bağlıdır. Bu nedenle açıklanabilir yapay zeka yöntemleri standart bir uygulama olarak teşvik edilmelidir (Mahayossanunt vd., 2023; Xie vd., 2021).
- Çok-kipli temel modeller ve dönüştürücü mimariler, alanın umut vadeden gelecek yönelimlerinden biri olarak öne çıkmaktadır. Dikkat mekanizmasına dayanan dönüştürücü mimari (Vaswani vd., 2017) uzun menzilli zamansal bağımlılıkları ve kipler arası etkileşimleri tek bir çatıda modelleyebildiğinden ses, görüntü ve metin ortak bir temsil uzayında birleştiren çok-kipli dönüştürücüler (multimodal transformers) için doğal bir zemin oluşturur (Xu vd., 2023). Etiketsiz büyük veride önceden eğitilip klinik göreve uyarlanan temel modellerle birleştiğinde bu mimariler, veri kısıtı sorununu azaltma ve tek bir modelle birden çok ruhsal bozukluğu ve klinik aşamayı kapsama potansiyeli taşır (Bommasani vd., 2021; Moor vd., 2023). Ancak bu yönelimin yukarıda

vurgulanan yanlılık, hesaplama maliyeti ve açıklanabilirlik koşullarıyla birlikte ilerlemesi gerekir.

Sonuç olarak, bu alandaki bir sonraki büyük atılım yalnızca model doğruluğunu artırmaya değil, araştırma odağını daha dengeli kılmaya, metodolojik ve etik güçlüklere yaratıcı çözümler üretmeye ve en önemlisi, klinik ihtiyaçları doğrudan karşılayan güvenilir ve yorumlanabilir sistemler inşa etmeye dayanmalıdır. Bu dönüşüm yalnızca teknik bir ilerleme değil, mühendislik, klinik bilimler ve etik arasında sürdürülebilir bir iş birliği gerektiren disiplinler arası bir olgunlaşma sürecidir. Alanın kendi başarısını ölçme biçimini doğruluk odaklı ölçütlerden klinik fayda odaklı ölçütlere kaydırması, bu olgunlaşmanın belirleyici göstergesi olabilir. Nihayetinde işitsel ve görsel verilerden elde edilen nesnel biyobelirteçlerin değeri laboratuvar başarımıyla değil, gerçek hastaların tanı, tedavi ve izlem yolculuğuna kattığı somut iyileşmeyle ölçülebilir.

Kaynakça

Adcock, A., & Parkin, L. (2016). Report from the independent Mental Health Taskforce to the NHS in England. House of Commons Library.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., … Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Das, A. K., & Naskar, R. (2024). A deep learning model for depression detection based on MFCC and CNN-generated spectrogram features. Biomedical Signal Processing and Control, 90, 105898. https://doi.org/10.1016/j.bspc.2023.105898

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP, a collaborative voice analysis repository for speech technologies. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 960–964).

de Melo, W. C., Granger, E., & Hadid, A. (2020). A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Transactions on Affective Computing, 13(3), 1581–1592. https://doi.org/10.1109/TAFFC.2020.3021755

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. Annual Review of Clinical Psychology, 14, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019. The Lancet Psychiatry, 9(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Gilanie, G., Asghar, M., Qamar, A. M., Ullah, H., Khan, R. U., Aslam, N., & Khan, I. U. (2022). An automated and real-time approach of depression detection from facial micro-expressions. Computers, Materials & Continua, 73(2). https://doi.org/10.32604/cmc.2022.028229

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., … Marsella, S. (2014). The Distress Analysis Interview Corpus of human and computer interviews (DAIC-WOZ). In Proceedings of LREC (Vol. 14, pp. 3123–3128).

Hall, N. T., Hallquist, M. N., Martin, E. A., Lian, W., Jonas, K. G., & Kotov, R. (2024). Automating the analysis of facial emotion expression dynamics: A computational framework and application in psychotic disorders. Proceedings of the National Academy of Sciences, 121(14), e2313665121. https://doi.org/10.1073/pnas.2313665121

He, L., Guo, C., Tiwari, P., Su, R., Pandey, H. M., & Dang, W. (2022). DepNet: An automated industrial intelligent system using deep learning for video-based depression analysis. International Journal of Intelligent Systems, 37(7), 3815–3835. https://doi.org/10.1002/int.22704

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., … Wang, Z. (2022). Deep learning for depression recognition with audiovisual cues: A review. Information Fusion, 80, 56–86. https://doi.org/10.1016/j.inffus.2021.10.012

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3), 404–413. https://doi.org/10.1038/nn.4238

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Kim, A. Y., Jang, E. H., Lee, S.-H., Choi, K.-Y., Park, J. G., & Shin, H.-C. (2023). Automatic depression detection using smartphone-based speech signals: Deep CNN approach. Journal of Medical Internet Research, 25, e34474. https://doi.org/10.2196/34474

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

Liu, X., Shen, H., Li, H., Tao, Y., & Yang, M. (2024). Multimodal depression detection based on self-attention network with facial expression and pupil. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2024.3405949

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116. https://doi.org/10.1002/lio2.354

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Mahayossanunt, Y., Nupairoj, N., Hemrungrojn, S., & Vateekul, P. (2023). Explainable depression detection based on facial expression using LSTM on attentional intermediate feature fusion with label smoothing. Sensors, 23(23), 9402. https://doi.org/10.3390/s23239402

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In IEEE International Conference on Data Engineering Workshops (ICDEW) (p. 8). https://doi.org/10.1109/ICDEW.2006.145

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. Trends in Cognitive Sciences, 16(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Monteith, S., Glenn, T., Geddes, J., & Bauer, M. (2015). Big data are coming to psychiatry: A general introduction. International Journal of Bipolar Disorders, 3(1), 21. https://doi.org/10.1186/s40345-015-0038-9

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., & Othmani, A. (2020). AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. Machine Learning with Applications, 2, 100005. https://doi.org/10.1016/j.mlwa.2020.100005

Othmani, A., Zeghina, A.-O., & Muzammel, M. (2022). A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. Computer Methods and Programs in Biomedicine, 226, 107132. https://doi.org/10.1016/j.cmpb.2022.107132

Prabhu, S., Mittal, H., Varagani, R., Jha, S., & Singh, S. (2022). Harnessing emotions for depression detection. Pattern Analysis and Applications, 25(3), 537–547. https://doi.org/10.1007/s10044-021-01020-9

Santomauro, D. F., Mantilla Herrera, A. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., … Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020

due to the COVID-19 pandemic. The Lancet, 398(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: A review. Briefings in Bioinformatics, 23(2), bbab569. https://doi.org/10.1093/bib/bbab569

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Uddin, M. A., Joolee, J. B., & Sohn, K.-A. (2022). Deep multi-modal network based automated depression severity estimation. IEEE Transactions on Affective Computing, 14(3), 2153–2167. https://doi.org/10.1109/TAFFC.2022.3179478

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., … Pantic, M. (2013). AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., … Pantic, M. (2016). AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Wang, H., Liu, Y., Zhen, X., & Tu, X. (2021). Depression speech recognition with a three-dimensional convolutional network. Frontiers in Human Neuroscience, 15, 713823. https://doi.org/10.3389/fnhum.2021.713823

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., & Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (pp. 11–18). https://doi.org/10.1145/2988257.2988263

World Health Organization. (2022a). COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. WHO News.

World Health Organization. (2022b). Mental health: Strengthening our response. WHO Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

World Health Organization. (2022c). Mental disorders. WHO Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Wu, W., Zhang, C., & Woodland, P. C. (2023). Self-supervised representations in speech-based depression detection. In ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10094910

Xie, W., Liang, L., Lu, Y., Wang, C., Shen, J., Luo, H., & Liu, X. (2021). Interpreting depression from question-wise long-term video recording of SDS evaluation. IEEE Journal of Biomedical and Health Informatics, 26(2), 865–875. https://doi.org/10.1109/JBHI.2021.3092628

Xu, N., Huo, H., Xu, J., Ma, L., & Wang, J. (2024). Automatic diagnosis of depression based on attention mechanism and feature pyramid model. PLoS One, 19(3), e0295051. https://doi.org/10.1371/journal.pone.0295051

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12113–12132. https://doi.org/10.1109/TPAMI.2023.3275156

Yang, W., Liu, J., Cao, P., Zhu, R., Wang, Y., Liu, J. K., … Zhang, X. (2023). Attention guided learnable time-domain filterbanks for speech depression detection. Neural Networks, 165, 135–149. https://doi.org/10.1016/j.neunet.2023.05.041

Zhang, X., Zhang, X., Chen, W., Li, C., & Yu, C. (2024). Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Scientific Reports, 14, 9543. https://doi.org/10.1038/s41598-024-60278-1

Zhou, X., Jin, K., Shang, Y., & Guo, G. (2020). Visually interpretable representation learning for depression recognition from facial images. IEEE Transactions on Affective Computing, 11(3), 542–552. https://doi.org/10.1109/TAFFC.2018.2828819

Chapter 7

A Stigmergy-Based Multi-Robot Search Strategy for Post-Earthquake Rubble Environments

Mehmet Dinçer Erbaş[1]

Abstract

Post-earthquake search and rescue operations require rapid exploration under uncertain conditions, high obstacle density and limited communication availability. In such environments multi-robot systems provide advantages in scalability and parallel exploration, yet effective coordination without centralized control remains a critical challenge. This study proposes a **Multi-Component Stigmergic Search (MCSS)** approach that extends conventional stigmergy-based coordination by integrating multiple environmental decision factors, including pheromone intensity, target-generated indirect signals, robot density and visitation history. The proposed approach was evaluated in a grid-based simulation environment representing rubble conditions and compared with a non-stigmergic exploration strategy and a fully random search method under different obstacle densities. Performance was assessed in terms of target discovery over time, cost per target, and average route length per target across repeated simulation runs. The results demonstrate that MCSS consistently achieves faster exploration, lower search cost, and shorter route lengths than the comparison approaches while maintaining stable performance under increasing environmental complexity. These findings suggest that combining stigmergic indirect communication with multi-component environmental guidance can improve the efficiency and robustness of autonomous multi-robot search strategies for disaster response scenarios.

1 Doç. Dr., Bolu Abant İzzet Baysal Üniversitesi, dincer.erbas@ibu.edu.tr, ORCID ID: 0000-0003-1762-0428

https://doi.org/10.58830/ozgur.pub1351.c5540

1. Introduction

Search and rescue operations conducted in post-earthquake rubble environments involve highly uncertain and physically hazardous conditions. In such scenarios, reaching victims within the shortest possible time is of critical importance. However, factors such as irregular debris structures, narrow passages, limited visibility, and the risk of secondary collapses significantly complicate conventional human-based search efforts. For this reason, increasing attention has been directed toward the use of robotic systems in search and rescue operations in recent years [1], [2], [3]. In particular, multi-robot systems offer significant advantages for search and rescue tasks due to their ability to perform parallel exploration, scale to large operational areas, and maintain robustness against individual robot failures [2].

Despite these advantages, achieving effective coordination among multiple robots without centralized control remains a major challenge [4]. A considerable portion of conventional coordination approaches relies on direct communication and global information sharing. However, in disaster environments, communication infrastructures may become unreliable or partially unavailable, thereby limiting the effectiveness of communication-dependent systems.

In this context, stigmergy provides an alternative coordination mechanism based on indirect communication through the environment [5]. In stigmergy-based approaches, robots can indirectly influence each other's behaviors through traces left in the environment, enabling collective search behaviors to emerge without centralized control. These characteristics make stigmergy a particularly promising approach for complex and communication-constrained search and rescue environments.

In this study, a stigmergy-based multi-robot search approach is investigated within a grid-based simulation environment representing post-earthquake rubble areas. In the proposed method, robot movement decisions are determined not only by pheromone intensity but also by multiple environmental factors, including target-generated signals, robot density, and visitation history. Furthermore, the performance of the proposed approach is comparatively evaluated against a non-stigmergic exploration strategy and a fully random search method. Within the scope of the study, performance metrics such as target discovery rate, cost per target, and route length are analyzed under different obstacle densities. The proposed framework also extends an earlier prototype simulation environment and stigmergy-based search structure previously investigated in a master's thesis study [6].

From a broader computational intelligence perspective, the proposed framework can also be interpreted within the context of emerging adaptive decision-making paradigms in artificial intelligence. Rather than relying on centralized planning or explicit communication, the proposed approach utilizes distributed environmental information and simple local interaction rules to generate collective exploration behavior. In this respect, the study reflects key principles of computational intelligence, including decentralized optimization, adaptive information processing, and environment-driven decision mechanisms. Furthermore, the integration of multiple environmental decision factors into a unified search strategy aligns with contemporary perspectives that emphasize scalable intelligent systems capable of extracting useful behavioral patterns from large and dynamically changing information spaces.

The main contributions of this study can be summarized as follows:

- A Multi-Component Stigmergic Search (MCSS) approach is proposed for multi-robot search in post-earthquake rubble environments by extending conventional stigmergy-based coordination with additional environmental decision components.
- The proposed decision mechanism integrates multiple environmental factors, including pheromone intensity, target-generated indirect signals, robot density, and visitation history, within a unified probabilistic movement-selection framework.
- A comparative evaluation framework is established by comparing the proposed MCSS approach against both a non-stigmergic exploration strategy and a fully random search method, enabling the isolated assessment of the contribution of stigmergic coordination.
- The proposed approach is evaluated under different obstacle-density scenarios using multiple performance indicators, including target discovery speed, cost per target, and average route length per target.
- The study extends an earlier prototype simulation framework through redesigned search dynamics, expanded environmental decision components, and a more comprehensive performance evaluation methodology for disaster-oriented multi-robot exploration.

2. Related Work

Structural collapse environments caused by disasters such as earthquakes, floods, explosions, and similar catastrophic events represent some of the most challenging operational settings for search and rescue missions. In such environments, reaching victims rapidly is critical for increasing survival

probability. However, rubble areas pose significant dangers to human rescue teams due to narrow passages, irregular obstacle distributions, limited visibility, dust, toxic gases, extreme temperatures, and the risk of secondary collapses. For this reason, the use of robotic systems as supportive tools in post-disaster search and rescue operations has been extensively investigated for many years. Studies in the field of Robotic Urban Search and Rescue (RUSAR) have demonstrated that robots can effectively be employed for exploration and information gathering in hazardous, inaccessible, or high-risk environments [1], [7]. In their review of urban search and rescue robots from a control perspective, Liu and Nejat identified mobility, sensing, mapping, autonomy, and human–robot interaction as the major research challenges in this domain [1]. Similarly, Drew emphasized that multi-robot systems provide substantial advantages for search and rescue applications in terms of scalability, robustness, and parallel exploration capabilities [2].

Teleoperation played a central role in early search and rescue robotic applications [1]. In teleoperated systems, robots are remotely controlled by human operators based on data obtained from onboard cameras and sensors [8]. Although this approach has demonstrated practical applicability in real disaster environments, it also presents several limitations. Restricted operator field of view, insufficient situational awareness, communication delays, and complex rubble structures may cause robots to become trapped or lose orientation. Experiences gained from robot-assisted search and rescue efforts following the September 11 attacks further highlighted the critical importance of human–robot interaction, situational awareness, and communication reliability in disaster operations [7]. Consequently, later studies focused not on completely abandoning teleoperation, but rather on increasing the autonomy level of robots and reducing operator workload [9], [10]. Nevertheless, single and operator-controlled robotic systems remain limited in their ability to rapidly scan large areas, simultaneously search for multiple targets, and maintain robustness against hardware failures.

These limitations have increased interest in the use of multi-robot systems for search and rescue problems. In multi-robot systems, tasks are distributed among multiple relatively simple robots instead of relying on a single highly complex platform [2]. This structure enables parallel exploration of the search area and improves system resilience against individual robot failures. Furthermore, increasing the number of robots allows the system to scale more effectively to larger or more complex operational environments. These characteristics make multi-robot systems particularly advantageous in uncertain, fragmented, and communication-constrained environments such as rubble fields [2]. However, one of the fundamental challenges in multi-

robot systems is achieving effective coordination without dependence on a centralized control unit [4]. Such coordination can be achieved through various mechanisms, including direct communication, global information sharing, or indirect communication through environmental traces.

Swarm intelligence approaches are widely employed in multi-robot exploration and search problems [11], [12]. Swarm intelligence is based on the emergence of complex collective behaviors from the interactions of numerous individuals operating according to relatively simple rules. Among the most well-known methods in this field are Particle Swarm Optimization and Ant Colony Optimization [13], [14]. Particle Swarm Optimization (PSO) is an optimization approach in which individuals update their movement directions using both their own experiences and the best experience of the group. Although this structure can be adapted to multi-robot search problems, it generally requires global information sharing or direct inter-robot communication. In rubble environments, however, the reliability of direct wireless communication may deteriorate due to concrete debris, metallic structural elements, irregular geometries, and signal attenuation. Consequently, methods that heavily depend on direct messaging and centralized information sharing may have limited applicability in disaster scenarios.

In this context, stigmergy offers an important alternative coordination mechanism for multi-robot search systems [5]. Stigmergy refers to indirect communication among individuals through traces left in the environment. This mechanism is commonly observed in nature, particularly in ant and termite colonies. Individuals deposit pheromone-like traces in the environment, while the behaviors of other individuals are influenced by the intensity and distribution of these traces. Theraulaz and Bonabeau described stigmergy as one of the fundamental mechanisms underlying self-organizing behaviors observed in social insects [15]. This structure enables the emergence of collective behavior without centralized control. From a robotics perspective, stigmergy is particularly attractive because it reduces communication overhead while allowing the environment itself to function as a form of shared memory.

The stigmergic approach also constitutes the foundation of the Ant Colony Optimization algorithm. Systematically formalized by Dorigo and colleagues, Ant Colony Optimization (ACO) is a swarm intelligence method in which artificial ants reinforce promising solutions over time by depositing pheromone trails throughout the solution space [13]. In this algorithm, pheromone accumulation increases the likelihood of reselecting previously successful paths, whereas pheromone evaporation reduces the influence of outdated or ineffective information. The combined use of these two mechanisms helps

establish a balance between exploration and exploitation. These characteristics have encouraged the application of stigmergy-based approaches to robotic problems such as target search, path planning, and area exploration.

In robotics literature, stigmergy is generally modeled through the concept of virtual pheromones rather than physical chemical substances. The "pheromone robotics" study conducted by Payton and colleagues is considered a pioneering work demonstrating that robot swarms can achieve coordination through virtual pheromone messages without centralized control [16]. In this approach, pheromones are represented not as chemical materials, but as numerical information maintained either by robots or within an environmental representation. The concept of virtual pheromones enables indirect coordination among robots in tasks such as surveillance, exploration, hazard detection, and path finding. Therefore, it provides a suitable modeling framework for simulation-based studies and digital map-based robotic applications that do not require physical pheromone emission.

In stigmergy-based robotic systems, pheromones can be defined as either attractive or repulsive. Attractive pheromones encourage robots to revisit successful paths or regions close to targets, whereas repulsive pheromones can discourage movement toward previously visited or potentially unproductive areas. Fossum and colleagues investigated the use of repulsive pheromones for efficient swarm robotic search in unknown environments and demonstrated that this approach can help reduce unnecessary revisits [17]. Similarly, Hamann and Wörn showed that swarm robotic search behaviors based on virtual pheromones can be modeled analytically and spatially [18]. These studies indicate that pheromone-based indirect communication is not merely a biologically inspired analogy, but also a computational mechanism that can be utilized to model and regulate the collective behaviors of robot swarms[16-19].

Nevertheless, the literature also demonstrates that stigmergy-based approaches do not always guarantee superior performance under all conditions. Hunt and colleagues examined the limitations of pheromone-based stigmergy in high-density robot swarms and reported that, under certain conditions, simple stigmergic avoidance behaviors may lose their advantage compared to random walk strategies [20]. This finding suggests that parameters such as robot density, environment size, obstacle distribution, pheromone evaporation rate, and target distribution must be carefully considered in stigmergy-based systems. Therefore, stigmergy should not be regarded as a standalone solution, but rather as a coordination mechanism that must be integrated with appropriate problem formulations, carefully selected parameters, and additional decision-making components.

From a search and rescue perspective, one of the most significant advantages of stigmergy-based approaches is the reduction of direct communication requirements. In rubble environments, wireless communication may become unreliable, and maintaining continuous connectivity among robots can be difficult. Under such conditions, decision-making mechanisms based on environmental traces or virtual environmental memory may provide a more robust coordination framework. Tang and colleagues proposed a stigmergy-based strategy for dynamic target search and tracking using swarm robots and demonstrated that a vector-based pheromone model can effectively support search and tracking behaviors [21]. Such studies reveal that stigmergy is not limited to static target search problems, but can also function as a flexible coordination mechanism under dynamic targets and changing environmental conditions.

A review of the existing literature reveals three major trends in search and rescue and multi-robot exploration research. The first trend is the transition from teleoperated and semi-autonomous systems toward more autonomous multi-robot structures [2], [9-10], [22-23]. The second trend is the shift from coordination mechanisms based on direct communication toward environment-mediated indirect communication approaches [4-5], [15]. The third trend involves extending simple pheromone-based guidance mechanisms with additional decision-making components such as visitation history, robot density, target signals, and environmental costs [17], [20-21], [24]. These trends highlight the increasing importance of multi-component decision mechanisms for complex environments characterized by uncertainty and high obstacle density, such as post-earthquake rubble search scenarios.

The simulation-based approach presented in this study can be positioned within this line of research. In the proposed method, robot movement decisions are not based solely on random selection or a single pheromone component. Instead, pheromone intensity, target-generated indirect signals, robot density, and cell visitation history are jointly considered within the decision-making process. This structure extends conventional pheromone-based search behavior by incorporating a more balanced and adaptive exploration mechanism. Furthermore, the study evaluates not only the performance of the stigmergy-based approach, but also compares it with a version of the same decision-making mechanism in which stigmergic components are disabled, as well as with a fully random search strategy. This comparative framework enables the contribution of stigmergic indirect communication to be analyzed separately against visitation-history-based exploration behavior and purely random movement strategies.

While many studies in the literature evaluate success primarily in terms of target acquisition or area coverage, the present study jointly considers target discovery speed, cost per target, and route length per target. This choice is particularly important in the context of search and rescue operations, since practical effectiveness depends not only on whether a target is found, but also on how quickly, efficiently, and with how little movement cost the target can be reached. In addition, the use of different obstacle densities allows the proposed method to be evaluated not only in relatively simple environments, but also in terms of its robustness under increasing environmental complexity. From this perspective, the present study extends existing stigmergy-based multi-robot search approaches within the specific context of post-earthquake rubble search problems at the simulation level. The primary contribution of the study is the investigation of a communication-independent, multi-component, probabilistic decision-making mechanism under different obstacle densities, together with a comparative evaluation of its performance against both random search and a non-stigmergic exploration strategy. A preliminary version of the simulation framework used in this study was previously explored in a master's thesis focusing on an ant-system-based control approach for rubble search problems [6]. In that earlier work, a basic stigmergy-based search structure and a limited set of environmental decision components were investigated. The present study substantially redesigns and extends this framework through the introduction of the proposed Multi-Component Stigmergic Search (MCSS) algorithm, additional environmental decision parameters, obstacle-density-based evaluations, expanded performance metrics, and a more comprehensive comparative analysis framework.

3. Multi-Component Stigmergic Search Algorithm

3.1 Problem Definition and Environment

Studies in the literature have demonstrated that stigmergy-based approaches can produce effective results in multi-robot exploration and search problems. However, in scenarios involving high obstacle density and severe mobility constraints, such as post-earthquake rubble environments, simple pheromone-based guidance mechanisms alone may not be sufficient. In particular, jointly considering additional components such as robot density, visitation history, and target-related environmental signals may contribute to the emergence of more balanced and efficient search behaviors. For this reason, this study investigates a simulation-based multi-robot search approach that extends stigmergy-based indirect communication with a multi-component decision-making structure. This approach will hereafter be referred to as the **Multi-Component Stigmergic**

Search (MCSS) algorithm. The simulation environment employed in this study was developed based on an earlier prototype framework introduced in a previous master's thesis study [6]. However, the current work significantly extends the original structure through redesigned search dynamics, multi-component probabilistic movement selection, obstacle-density-based scenario generation, and additional performance evaluation criteria including search cost and route efficiency.

The problem addressed in this study is a multi-robot target search problem representing search and rescue operations conducted in post-earthquake rubble environments. The problem is defined as the rapid detection of targets (victims) located within a structurally complex environment containing obstacles, using autonomous robots. Within this framework, the robots are expected to exhibit collective search behavior without relying on a centralized control mechanism, instead coordinating through environmental information and indirect communication mechanisms.

The simulation environment is modeled as a two-dimensional discrete structure. The search area is represented using a grid-based environment composed of equally sized cells. This representation was selected to abstract the irregular, fragmented, and mobility-constrained characteristics of post-earthquake rubble fields. Each cell may contain free space, an obstacle, or a target. An example simulation environment is illustrated in Fig. 1. The figure presents the irregular obstacle distribution within the grid-based search area, together with the initial positions of the robots and the spatial placement of the targets. This structure visually demonstrates both the mobility constraints encountered by the robots and the complexity of the operational environment.

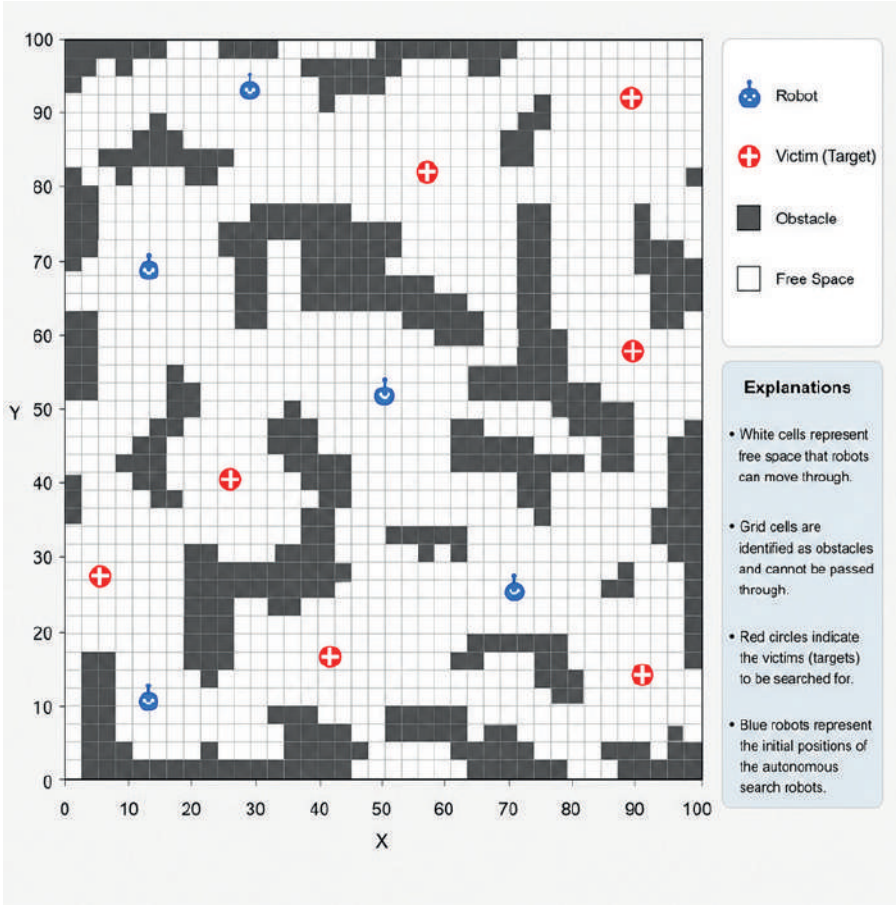


Fig 1. Example simulation environment containing obstacles. In the grid-based structure, white cells represent free spaces through which robots can move, while gray cells represent obstacles that cannot be traversed. Blue markers indicate the current positions of the robots, whereas red markers denote the locations of the search targets (victims).

Obstacles within the environment are modeled as non-traversable cells that restrict robot movement. These obstacles represent physical constraints such as rubble fragments, collapsed structural components, and narrow passageways commonly encountered in post-earthquake environments. The obstacle distribution was designed to increase environmental complexity, thereby preventing robots from moving freely and directly across the search area. This configuration enables the search problem to incorporate more realistic operational challenges.

Multiple targets are distributed throughout the simulation environment at different spatial locations. Rather than being directly observable by the

robots, these targets are represented through indirect signals that can be perceived within a limited influence range. This approach abstracts real-world search and rescue scenarios in which victims trapped beneath rubble cannot be directly seen, but may instead be detected through indirect indicators such as sound, movement, or similar signals. The robots operating within the environment are initially positioned at different starting locations. Robots move over discrete time steps and, at each step, are allowed to transition only between neighboring cells. Movement decisions are determined using local environmental information rather than a centralized control mechanism. Robots are not permitted to enter cells containing obstacles. This framework enables the investigation of both multi-robot coordination behavior and the effects of the stigmergy-based indirect communication mechanism employed in the MCSS approach under complex search conditions.

In the MCSS approach, robots leave virtual traces within the environment. Other robots perceive these traces and use them to shape their movement decisions. The influence of these environmental traces gradually decreases over time, thereby creating a dynamic information structure within the environment. This mechanism provides an important advantage in disaster scenarios where direct communication is unavailable or severely limited. The simulation process is repeated multiple times under different initial conditions. These repetitions help reduce the influence of randomness and enable more reliable performance evaluations. The defined problem structure reflects the primary challenges encountered in post-earthquake rubble search and rescue scenarios, including obstacle density, mobility constraints, uncertainty, and multi-target search. This framework therefore enables a comparative evaluation of different search strategies under complex operational conditions.

3.2 Stigmergy-Based Search Strategy

In this study, the **Multi-Component Stigmergic Search (MCSS)** approach is employed, in which robots navigate by utilizing environmental traces and indirect signals associated with targets. Within the MCSS framework, robots do not move purely randomly. Instead, the decision-making process incorporates both environmental traces and indirect target-related signals. The proposed approach is inspired by the fundamental principles of Ant Colony Optimization (ACO) [13]. However, the classical structure has been extended and adapted to suit the specific characteristics of the problem domain.

At each time step, every robot evaluates the neighboring cells that are reachable from its current position. Directions containing obstacles are eliminated, and the remaining feasible movement options are identified. For

each candidate cell, a preference weight is calculated. This weight is determined by the combination of four main components:

- Pheromone intensity
- Target signal (sound)
- Robot density
- Visitation history

The movement tendency of a robot from position i to a neighboring cell j is expressed as follows:

$$w_{i,j} = \left(1+\tau_{i,j}\right)^{\alpha}\left(1+S_{i,j}\right)^{\beta}\left(1+R_{i,j}\right)^{\gamma}\cdot\frac{1}{1+V_{i,j}} \tag{1}$$

In the above formulation:

$\tau_{i,j}$: pheromone intensity associated with the corresponding cell

$S_{i,j}$: intensity of the indirect signal originating from targets

$R_{i,j}$: term representing robot density

$V_{i,j}$: number of previous visits to the corresponding cell

α , β , γ : coefficients determining the influence of the respective components

This structure ensures that robots do not rely on a single source of information. Instead, multiple environmental factors are jointly evaluated to establish a more balanced decision-making mechanism.

Pheromone intensity enables robots to preferentially follow paths that were previously explored successfully. When a robot reaches a target, it deposits pheromones along the route it followed. This process can be expressed as follows:

$$\tau_{i,j} \leftarrow \left(1-\rho\right)\tau_{i,j} + \Delta\tau_{i,j} \tag{2}$$

Here, ρ represents the pheromone evaporation rate. This parameter controls the gradual reduction of pheromone intensity over time and prevents outdated information from becoming dominant within the system. The sound generated by targets is not directly observable. Instead, it is represented through a signal that can be perceived within a limited influence range. This signal guides robots toward potential target locations and supports the search process under conditions where direct target observation is not possible:

$$S_{i,j} = \frac{1}{d_{i,j} + \epsilon} \tag{3}$$

Here, $d_{i,j}$ denotes the distance between the corresponding cell and the target. The parameter ϵ is a small positive constant introduced to ensure numerical stability and to prevent division-by-zero conditions. Accordingly, the perceived signal intensity increases as the robot approaches the target. As a result, cells located closer to the target are assigned higher preference values.

The presence of a large number of robots within the same region may reduce search efficiency due to overcrowding and redundant exploration. For this reason, robot density is incorporated into the decision-making mechanism. Robot density is defined as follows:

$$R_{i,j} = \frac{N_{i,j}}{N_{\max}} \tag{4}$$

Here, $N_{i,j}$ represents the number of robots located within the corresponding cell or its surrounding neighborhood. The parameter $N_{\max}$ denotes the maximum robot density used for normalization purposes. This term reduces excessive clustering of robots within the same region and encourages a more balanced distribution across the search area.

The tendency of robots to repeatedly revisit frequently explored cells is intentionally reduced. For this purpose, a penalty term based on visitation history is incorporated into the preference weight. This effect is modeled as follows:

$$F_{i,j} = \frac{1}{1 + V_{i,j}} \tag{5}$$

Here, $V_{i,j}$ denotes the number of times the corresponding cell has previously been visited. As the visitation count increases, the value of this term decreases. Consequently, the probability of selecting frequently visited cells becomes lower. This mechanism limits unnecessary repeated exploration within the same regions and encourages the discovery of previously unexplored areas.

The movement decision is determined probabilistically according to the transition tendency values associated with traversable neighboring cells. The preference weights calculated for each neighboring cell are normalized to obtain a transition probability:

$$P_{i,j} = \frac{w_{i,j}}{\sum_{k \in \mathcal{N}_{\square}} w_{i,k}} \tag{6}$$

Here, $\mathcal{N}_{\square}$ denotes the set of neighboring cells that are reachable from the robot's current position. Through this structure, robots are more likely to select directions associated with higher weight values. However, the probability of selecting other feasible directions is not completely eliminated. As a result, the system maintains a balance between directed search behavior and exploratory behavior. The pseudocode of the proposed MCSS algorithm is presented in **Algorithm 1**.

Algorithm 1. Multi-Component Stigmergic Search (MCSS) Algorithm

1. **For each robot, determine the neighboring cells that are reachable from the current position.**
2. **Remove obstacle-containing or non-traversable cells from the set of candidate movements.**
3. **For each candidate cell, calculate the environmental attractiveness value by considering:**
   - **the pheromone intensity in the corresponding direction,**
   - **the influence of target-generated signals,**
   - **the robot density within the same region,**
   - **the reduced selection probability of previously visited cells.**
4. **Convert the attractiveness values of the candidate cells into transition probabilities.**
5. **Select the robot's next movement according to these probabilities.**
6. **Move the robot to the selected neighboring cell.**
7. **Update the robot's visitation information and traversed path.**
8. **If the robot reaches a target:**
   8.1 Mark the target as found.
   8.2 Retrieve the route followed by the robot while reaching the target.

8.3 Deposit pheromones along the transitions on this route.
8.4 Return the robot to its initial position.

9. **Apply pheromone evaporation at each time step by reducing pheromone values.**
10. **Repeat the process until all targets are found or the search procedure terminates.**

When a robot reaches a cell containing a target, the corresponding target is marked as found. The value associated with the target cell is then reset to zero, preventing the same target from being counted again in subsequent time steps. Within the MCSS approach, pheromone traces are reinforced along the route followed by the robot while reaching the target. This process enables successful search routes to be propagated throughout the environment. After reaching a target, the robot is returned to its initial position and reintroduced into the search process. This structure allows robots to continue operating continuously in environments containing multiple targets.

The random search strategy illustrates the behavior of the multi-robot system in the absence of environmental information and indirect coordination mechanisms. In contrast, the MCSS approach investigates how robots can exhibit more directed search behavior by utilizing environmental traces and indirect target-related signals. By comparing these two approaches, the effects of the stigmergic coordination mechanism on target discovery, search cost, and robot movement efficiency can be evaluated. In this way, the study examines how the stigmergy-based indirect communication mechanism discussed in the literature can contribute to search and rescue problems within a representative simulation environment.

3.3 Experimental Setup

To evaluate the performance of the MCSS approach, simulation experiments based on a multi-robot exploration problem were conducted. The MCSS approach was compared against two alternative strategies: (i) an exploration approach without a stigmergic coordination mechanism and (ii) a fully random search strategy. The experiments were performed in a two-dimensional discrete environment consisting of a $100 \times 100100 \times 100$ grid structure. Within this environment:

- 10 targets were randomly distributed throughout the search area,
- 12 robots were initially positioned at random locations,

- different numbers of obstacles were introduced to control environmental complexity.

At the beginning of each experiment, robots, targets, and obstacles were randomly distributed without overlap. At each time step, robots were allowed to move to one of the eight neighboring cells. Every movement decision was subject to obstacle checking, and the next position was selected only from feasible directions. Movements toward obstacle-containing cells were not permitted. If a robot had no valid movement option, it remained stationary for that time step.

When a robot reached a cell containing a target, the route followed by the robot to reach that target was recorded, and the robot was returned to its initial position. In the MCSS approach, robots deposit pheromone traces into the environment during this return process. As described in Section 3.2, the selection value for each possible movement is calculated probabilistically based on pheromone intensity, target-generated signals, robot density, and the visitation count of the corresponding cell. The resulting selection probabilities are normalized, and the next movement is determined through probabilistic sampling.

As previously noted, whenever a robot discovers a target, pheromone traces are reinforced along the traversed route. In addition, pheromone intensity is gradually reduced over time through an evaporation mechanism applied at each time step. The parameter values used in all experiments are presented in Table 1.

Table 1. Simulation Parameters

Parameter	**Value**
Environment size	100 x 100 cells
Number of robots	12
Number of targets	10
Number of obstacles	250 / 500 / 750
Maximum time steps	10,000
Number of experiment repetitions	100
Pheromone evaporation rate	0.02
σ (pheromone influence)	1
γ (target signal influence)	1
β (robot density influence)	1

The parameter values presented in Table 1 were selected to provide a controlled and computationally feasible evaluation environment while maintaining comparable influence among the decision components. The environment size (100 × 100 cells), number of robots (12), and number of targets (10) were chosen to create a sufficiently large search space that allows collective behaviors to emerge without introducing excessive computational cost. Obstacle densities of 250, 500, and 750 cells were used to represent progressively increasing environmental complexity and mobility constraints. The pheromone evaporation rate was set to 0.02 to preserve environmental memory while preventing outdated information from dominating the search process over long simulation periods. The influence coefficients of pheromone intensity ($\sigma = 1$), target-generated signals ($\gamma = 1$), and robot density ($\beta = 1$) were initialized with equal values to avoid introducing a priori preference toward any individual decision component and to allow the combined effect of the multi-component mechanism to be evaluated more transparently. These parameter values were determined through preliminary simulation trials aimed at obtaining stable search behavior and ensuring meaningful comparisons across different search strategies rather than performing exhaustive parameter optimization.

In this study, three different search approaches were compared. The first approach, MCSS, was described in detail in Section 3.2. The second approach, used for comparative purposes, was derived from the same decision-making framework with the stigmergic components disabled. In this configuration, the overall algorithmic structure was preserved, while the pheromone influence (σ), target-generated signal influence (γ), and robot density influence (β) parameters used during movement selection were set to zero. As a result, environmental traces and target-oriented guidance were removed, while the visitation-history component of the decision mechanism was retained. Consequently, robots exhibited an exploration-oriented behavior by preferentially moving toward less frequently visited cells. The third approach was a fully random search strategy. In this strategy, robots selected among feasible movements with equal probability and did not utilize any memory or guidance mechanism during the search process.

When these three approaches are evaluated together, the contribution of the stigmergic coordination mechanism can be analyzed comparatively against both visitation-history-based exploration behavior and a fully random movement strategy. The primary performance metrics considered in the experiments were the average number of targets found over time, robot movement cost per target, and route length per target. All results were obtained by averaging 100 independent simulation runs and are presented in the figures together with

95% confidence intervals. Although a dedicated component-wise ablation study was not originally designed, the comparative evaluation framework adopted in this study can also be interpreted as a functional ablation analysis of the proposed MCSS architecture. Specifically, comparison between the full MCSS configuration and the exploration strategy with disabled stigmergic components enables the isolated contribution of stigmergy-based indirect coordination to be evaluated while preserving the remaining exploration mechanism. Furthermore, comparison with the fully random search strategy provides an additional lower-bound reference in which all environmental guidance mechanisms are removed. This structure allows the incremental contribution of the proposed coordination mechanism to be analyzed without requiring a separate ablation protocol.

To evaluate whether the observed performance differences among the search strategies were statistically meaningful, additional statistical analyses were performed using the results obtained from 100 independent simulation runs for each experimental configuration. Prior to statistical comparison, normality assumptions were evaluated for the collected performance metrics and confirmed to be satisfied. Accordingly, pairwise comparisons among the search approaches were conducted using independent-samples t-tests. Since multiple pairwise comparisons were performed among the three evaluated approaches, p-values were adjusted using the Holm–Bonferroni correction procedure to control the family-wise error rate. Statistical significance was evaluated at $\alpha = 0.05$. The statistical analysis was conducted separately for each obstacle-density scenario and for each performance metric considered in the study.

4. Results

In this study, the MCSS approach was compared with the exploration approach in which stigmergic components were disabled and with the fully random search strategy using three different performance metrics: (i) the number of targets found over time, (ii) cost per target, and (iii) average route length per target. The experiments were conducted in three different environments containing 250, 500, and 750 obstacles, respectively, and each configuration was evaluated over 100 independent simulation runs. In all figures, average values are presented together with 95% confidence intervals.

4.1 Average Number of Targets Found Over Time

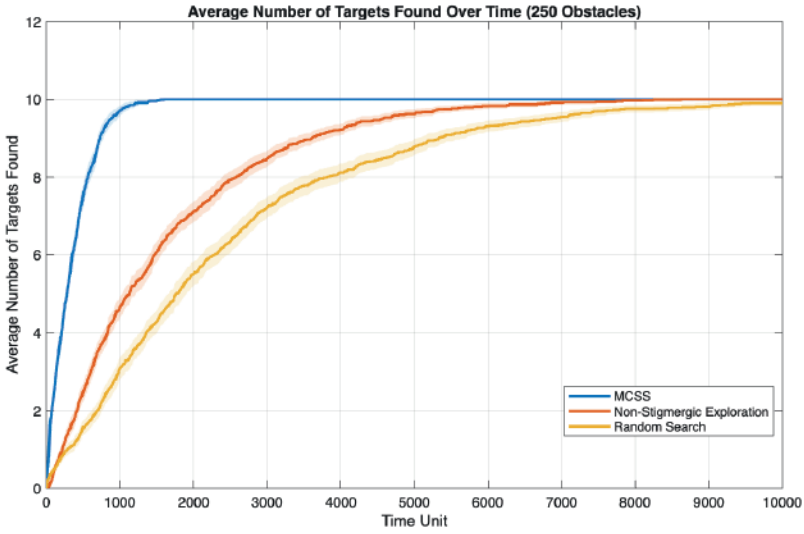
The results presented in Fig. 2 demonstrate that the three approaches exhibit clearly different exploration speeds and convergence behaviors. Across all obstacle densities, the MCSS approach achieved the fastest overall performance

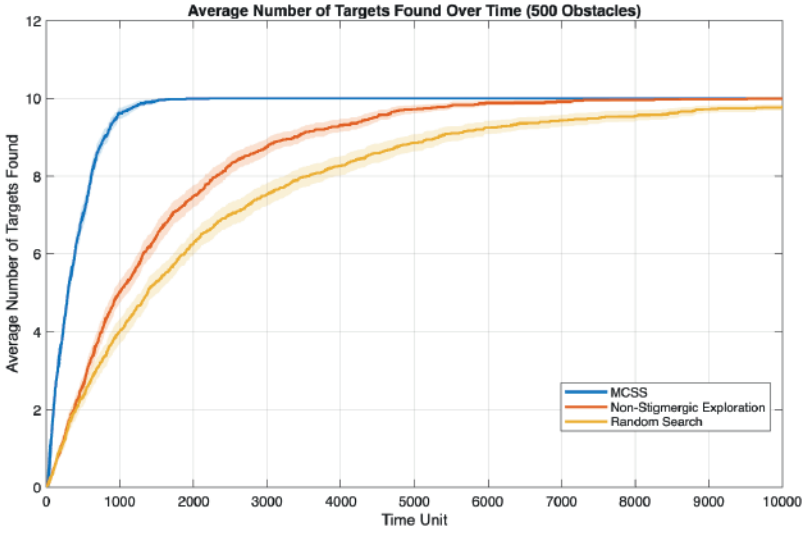
by a substantial margin. This approach was able to discover the majority of targets during the early stages of the simulation and reached the maximum number of targets within approximately 1000–1500 time steps.

In contrast, the exploration approach with disabled stigmergic components exhibited a slower but more gradual and stable increase in performance. Although this approach performed significantly better than the fully random search strategy, it was unable to achieve the early exploration advantage demonstrated by the MCSS approach. The random search strategy produced the lowest performance overall, requiring considerably longer times to reach targets and remaining notably behind the other approaches, particularly during the early stages of the simulations.

As the number of obstacles increased, a decrease in performance was observed for all approaches, as expected. However, the MCSS approach was the least affected by the increase in environmental complexity. Notably, even in the scenario containing 750 obstacles, the approach maintained its rapid early convergence behavior, indicating strong capabilities in navigation and indirect information sharing within complex environments. Furthermore, an examination of the confidence intervals reveals that the MCSS approach exhibited lower variance compared to the alternative methods. This finding suggests that the proposed approach provides not only fast but also stable and consistent performance across different simulation runs.

To further evaluate whether the observed differences were statistically meaningful, statistical comparisons were performed using pairwise independent-samples t-tests with Holm–Bonferroni correction. The results supported the visual trends presented in Fig. 2. Compared with the non-stigmergic exploration approach, the MCSS approach achieved statistically significant improvements in the average number of targets found during approximately the first 6000 time units (adjusted $p < 0.05$). After this stage, the performance difference gradually decreased as both approaches approached convergence. In contrast, the MCSS approach maintained statistically significant superiority over the random search strategy throughout the entire simulation period (adjusted $p < 0.05$), indicating that stigmergic coordination contributed substantially to faster and more effective exploration.


Average Number of Targets Found Over Time (250 Obstacles)
Average Number of Targets Found
Time Unit
MCSS
Non-Stigmergic Exploration
Random Search
0
2
4
6
8
10
12
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


Average Number of Targets Found Over Time (500 Obstacles)
Average Number of Targets Found
Time Unit
MCSS
Non-Stigmergic Exploration
Random Search
0
2
4
6
8
10
12
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

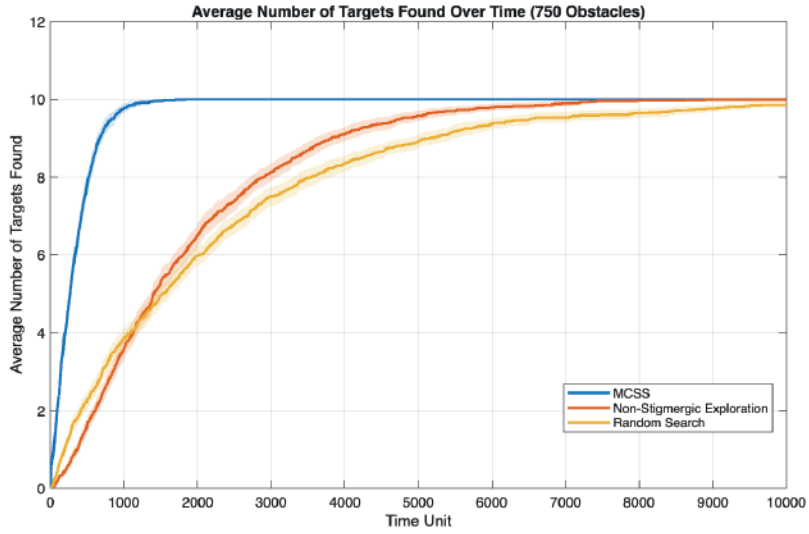


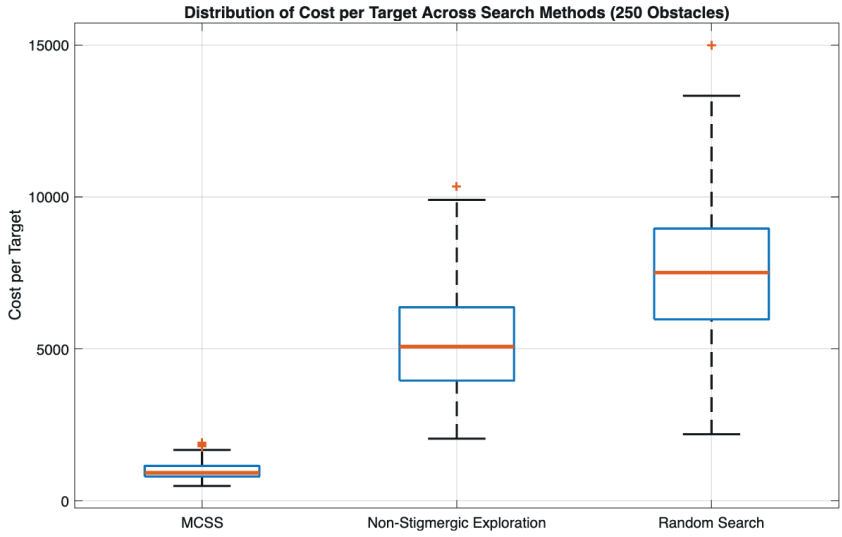
Fig. 2. Comparison of the average number of targets found over time for the three search methods under different obstacle densities (250, 500, and 750 obstacles). The curves represent the average values obtained from 100 independent simulation runs, while the shaded regions indicate the 95% confidence intervals. The Multi-Component Stigmergic Search (MCSS) approach exhibits faster convergence and earlier target discovery across all scenarios, the non-stigmergic exploration method demonstrates intermediate performance, and the random search strategy produces the lowest exploration speed.

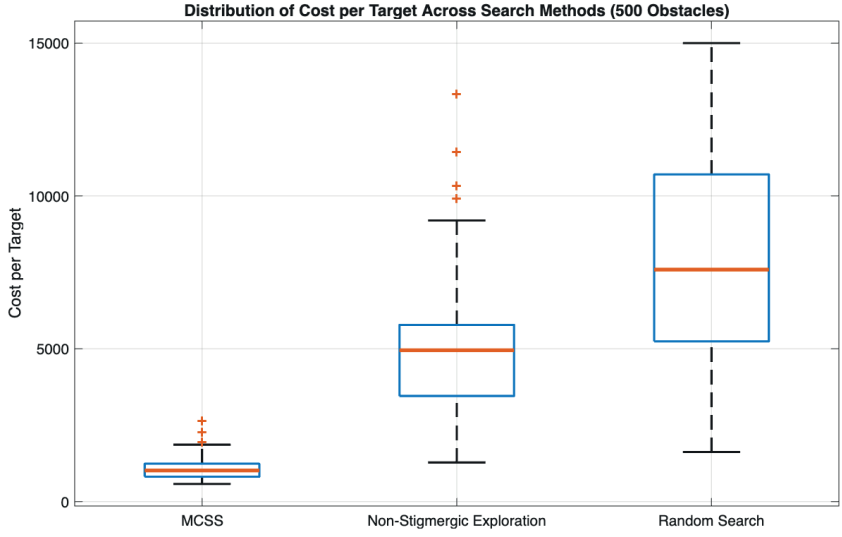
4.2 Cost per Target

The boxplots presented in Fig. 3 clearly reveal the differences among the approaches in terms of cost per target. Across all experimental environments, the MCSS approach achieved the lowest cost values. The combination of low median values and narrow distributions indicates that this approach provides both efficient and consistent performance. The exploration approach with disabled stigmergic components demonstrates a noticeable improvement compared to the random search strategy. By prioritizing less frequently visited cells, this approach increases exploration efficiency and thereby reduces search cost. However, because it lacks the indirect information-sharing mechanism provided by stigmergy, its cost values remain higher than those of the MCSS approach, particularly in complex environments. The random search strategy exhibits the highest cost values overall. In this approach, the routes followed to reach targets are largely inefficient, and unnecessary revisits occur frequently. Examination of the boxplots further shows that this strategy has a considerably wider distribution and contains a large number of outliers. This observation indicates that random search produces highly unpredictable and unstable performance. As the number of obstacles increases, cost values rise for all approaches. However, the increase remains relatively limited in the MCSS

approach, whereas it becomes substantially more pronounced in the alternative methods. These results suggest that the MCSS approach offers a more scalable structure for operation in complex environments.

To assess whether the observed differences in search efficiency were statistically meaningful, pairwise comparisons were conducted using independent-samples t-tests with Holm–Bonferroni correction. The statistical analysis confirmed that the MCSS approach achieved significantly lower cost-per-target values than both the non-stigmergic exploration approach and the random search strategy across all evaluated obstacle-density scenarios (adjusted $p < 0.05$). In contrast, the difference between the non-stigmergic exploration approach and random search gradually decreased as environmental complexity increased and became statistically non-significant under higher obstacle-density conditions. This finding suggests that the visitation-history-based exploration mechanism alone provides limited improvements in search efficiency and that the indirect coordination enabled by stigmergic information becomes increasingly important in more constrained environments.


Distribution of Cost per Target Across Search Methods (250 Obstacles)
Cost per Target
15000
10000
5000
0
MCSS
Non-Stigmergic Exploration
Random Search


Distribution of Cost per Target Across Search Methods (500 Obstacles)
Cost per Target
15000
10000
5000
0
MCSS
Non-Stigmergic Exploration
Random Search

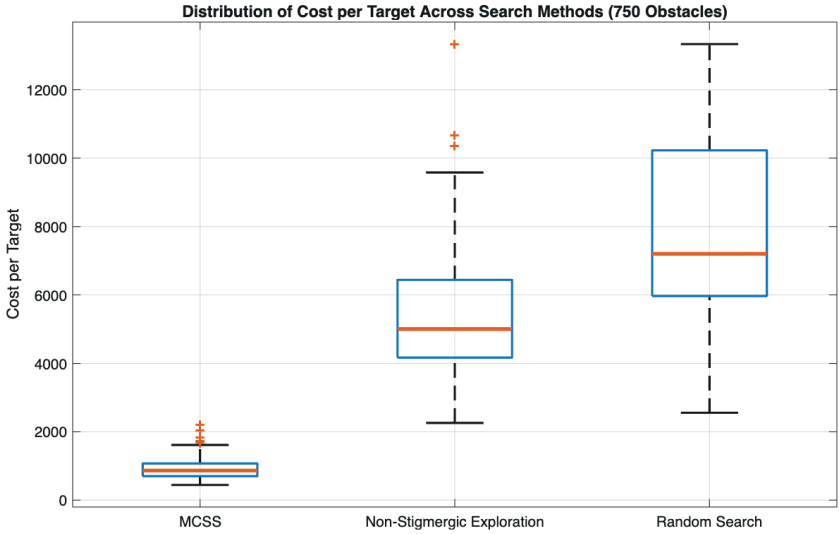


Fig. 3. Comparison of the cost-per-target distributions for the three search approaches under different obstacle densities (250, 500, and 750 obstacles). The boxplots illustrate the median values, interquartile ranges, and outliers obtained from 100 independent simulation runs. The Multi-Component Stigmergic Search (MCSS) approach achieves the lowest cost and the narrowest distribution across all scenarios, indicating the most efficient and stable performance. The exploration approach with disabled stigmergic components produces lower costs than the random search strategy, but still remains behind the MCSS approach. The random search strategy exhibits the highest costs and the widest variance, demonstrating the lowest overall efficiency.

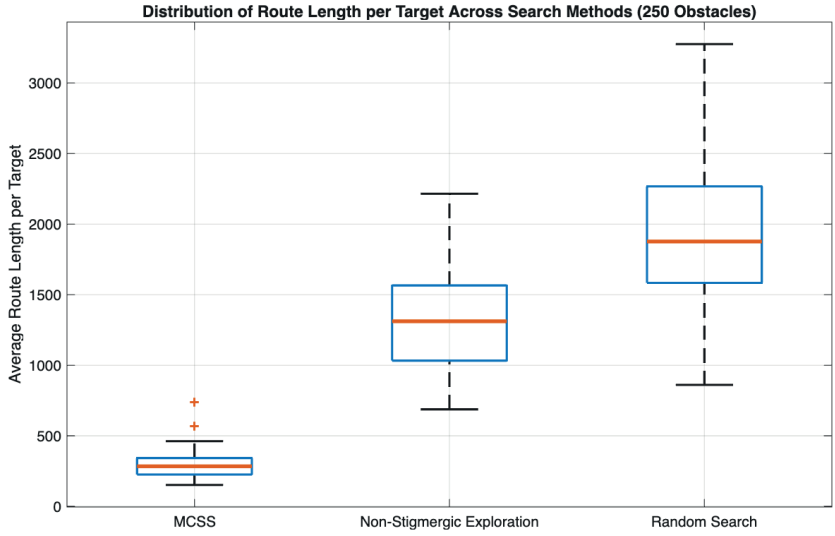
4.3 Average Route Length per Target

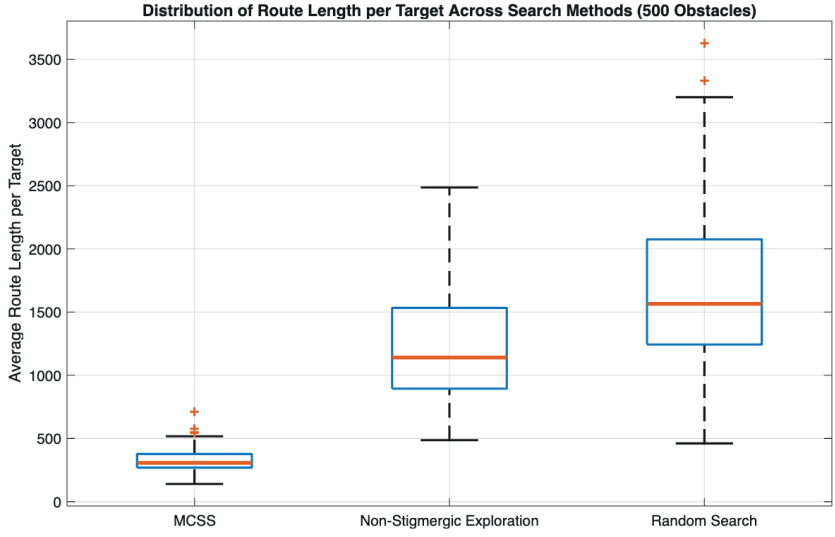
The results presented in Fig. 4 demonstrate the effectiveness of the routes followed by the approaches in reaching targets. The findings clearly show that the MCSS approach produces the shortest route lengths among all methods. Through the use of environmental traces, robots in the MCSS framework avoid repeatedly exploring previously visited regions and instead follow more direct and efficient paths. This behavior reduces unnecessary wandering and minimizes route lengths. Examination of the boxplots further indicates that the MCSS approach exhibits both low median values and a narrow distribution.

The exploration approach with disabled stigmergic components performs better than the random search strategy in terms of route length. The strategy of favoring less frequently visited cells provides a certain degree of exploration efficiency and contributes to shorter routes. However, because this approach does not benefit from the collective information-sharing mechanism provided by stigmergy, its route optimization capability remains limited compared to the MCSS approach.

The random search strategy generates the longest route lengths overall. In this approach, the paths followed to reach targets are largely indirect and contain frequent repetitive movements. Moreover, the wide distribution observed in the boxplots indicates unstable and inconsistent behavior. As the number of obstacles increases, route lengths rise for all approaches. Nevertheless, the MCSS approach is the method least affected by this increase. This finding suggests that the proposed approach maintains effective path-planning capability even in environments with high obstacle density.

To evaluate whether the observed differences in route efficiency were statistically meaningful, pairwise comparisons were conducted using independent-samples t-tests with Holm–Bonferroni correction. The statistical analysis confirmed that the MCSS approach achieved significantly shorter average route lengths per target than both the non-stigmergic exploration approach and the random search strategy across all evaluated obstacle-density scenarios (adjusted $p < 0.05$). In contrast, the difference between the non-stigmergic exploration approach and random search progressively diminished as obstacle density increased and became statistically non-significant under more complex environmental conditions. These findings indicate that reducing route inefficiency in constrained search environments depends not only on avoiding previously visited regions but also on the indirect coordination and collective guidance provided by the stigmergic mechanism.


Distribution of Route Length per Target Across Search Methods (250 Obstacles)
Average Route Length per Target
0
500
1000
1500
2000
2500
3000
MCSS
Non-Stigmergic Exploration
Random Search


Distribution of Route Length per Target Across Search Methods (500 Obstacles)
Average Route Length per Target
0
500
1000
1500
2000
2500
3000
3500
MCSS
Non-Stigmergic Exploration
Random Search

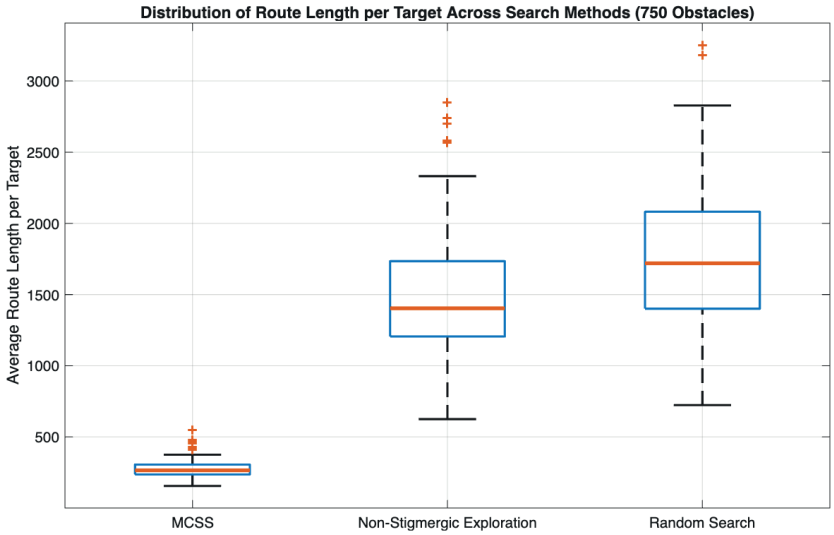


Fig. 4. Comparison of the average route-length distributions per target for the three search approaches under different obstacle densities (250, 500, and 750 obstacles). The boxplots present the median values, interquartile ranges, and outliers obtained from 100 independent simulation runs. The Multi-Component Stigmergic Search (MCSS) approach achieves the shortest route lengths and the narrowest distributions across all scenarios, demonstrating the most efficient path-planning performance. The exploration approach with disabled stigmergic components produces shorter routes than the random search strategy, but still remains behind the MCSS approach. The random search strategy exhibits the longest and most variable route lengths, resulting in the lowest overall performance.

When all results are evaluated collectively, both the graphical observations and the statistical analyses consistently indicate that the MCSS approach provides clear advantages in three major aspects:

- Faster exploration performance (early convergence)
- Lower search cost
- Shorter and more efficient routes

Pairwise comparisons performed using independent-samples t-tests with Holm–Bonferroni correction confirmed that the MCSS approach achieved statistically significant improvements over random search across all evaluated performance metrics and obstacle-density scenarios. Furthermore, compared with the non-stigmergic exploration approach, the MCSS approach also demonstrated statistically significant advantages in target discovery performance, cost per target, and average route length under the evaluated conditions.

The non-stigmergic exploration approach produced measurable improvements over random search, particularly in terms of exploration

efficiency during the earlier stages of the search process. However, the statistical analysis further revealed that these advantages diminished as obstacle density increased and became non-significant for some efficiency-related metrics under more complex environmental conditions. This finding suggests that visitation-history-based guidance alone contributes positively to exploration behavior but remains insufficient to maintain robust search efficiency in constrained environments.

Overall, the combined experimental and statistical findings demonstrate that the MCSS approach offers an effective and scalable solution for multi-robot exploration problems and provides particularly strong advantages in environments with high obstacle density through stigmergy-based indirect coordination.

5. Conclusion

In this study, stigmergy-based multi-robot search approaches developed for post-earthquake search and rescue problems were investigated, and the performance of the proposed Multi-Component Stigmergic Search (MCSS) approach was evaluated in a simulation environment. Within the scope of the study, the effects of extending conventional stigmergy-based guidance mechanisms with additional environmental components such as robot density, visitation history, and target-generated indirect signals were examined in the context of multi-robot exploration behavior. For this purpose, the MCSS approach was compared against an exploration approach with disabled stigmergic components and a fully random search strategy.

The obtained results demonstrate that the MCSS approach provides clear advantages over the alternative methods, particularly in terms of early exploration performance, low movement cost, and short route generation. Analysis of the number of targets found over time showed that the MCSS approach achieved faster convergence and reached targets more rapidly across all obstacle densities. Furthermore, the cost-per-target and route-length results revealed that stigmergic indirect communication through environmental traces not only accelerates exploration but also improves the efficiency of robot movements. The limited performance degradation observed in high-obstacle-density scenarios further indicates that the proposed approach offers a scalable and robust structure for operation in complex environments.

The exploration approach with disabled stigmergic components produced better results than the random search strategy. This finding suggests that the visitation-history-based guidance mechanism contributes positively to the exploration process. However, due to the absence of indirect information

sharing through environmental traces, this approach could not fully achieve the collective guidance advantages provided by the MCSS framework. The random search strategy, on the other hand, demonstrated the lowest performance across all evaluation metrics and exhibited highly inefficient behavior, particularly in environments with high obstacle density.

From a practical performance perspective, the proposed MCSS approach demonstrated substantial quantitative gains across different environmental conditions. Relative to the non-stigmergic exploration approach, MCSS reduced the average cost per target by approximately 78–83% and shortened the average route length per target by approximately 73–82%. Compared with random search, the improvements became even more pronounced, corresponding to approximately 87–88% lower search cost and 80–85% shorter average route lengths. These improvements were further supported by statistical analysis, which confirmed statistically significant differences with consistently large effect sizes across the evaluated efficiency-related metrics. Collectively, these findings indicate that the proposed stigmergic coordination mechanism provides not only statistically significant improvements but also practically meaningful gains with large effect sizes, indicating that the observed performance differences are unlikely to be attributable to random variation alone.

Overall, the obtained findings indicate that stigmergy-based indirect communication mechanisms can be effectively utilized in multi-robot search and rescue problems. The fact that the MCSS approach does not require centralized control, is not dependent on direct communication, and supports collective exploration behavior makes it particularly attractive for disaster scenarios in which communication infrastructure is limited or unreliable. In addition, the modular structure of the approach allows different environmental components to be incorporated into the decision-making mechanism, thereby providing a flexible foundation for the development of more advanced multi-robot systems in future studies. From a broader computational intelligence perspective, the findings further illustrate how adaptive collective behavior can emerge through distributed information processing and environment-driven decision mechanisms, supporting the development of scalable intelligent systems for complex search environments.

Nevertheless, the study has several limitations. The simulation environment was modeled as a two-dimensional discrete structure and therefore does not fully capture the physical complexity of real-world disaster environments. Robot movements were evaluated under idealized assumptions, while factors such as sensor errors, mechanical failures, communication delays, and dynamic

environmental changes were not included in the model. Furthermore, the virtual stigmergy mechanism employed in this study does not directly incorporate perception and environmental interaction challenges that may arise in real physical environments.

In addition, several practical constraints commonly encountered in real-world robotic search operations were not explicitly modeled in the present study. The proposed framework assumes reliable localization, synchronized robot operation, and ideal environmental state updates, whereas real deployments may involve positioning uncertainty, asynchronous robot behavior, limited onboard computation, sensor noise, actuator inaccuracies, and intermittent communication availability. Furthermore, the virtual stigmergic representation used in the simulation assumes consistent environmental memory and instantaneous information propagation, while real implementations may require distributed map synchronization and introduce additional latency and uncertainty. These factors may influence both exploration efficiency and coordination quality and should therefore be incorporated into future validation studies involving more realistic robotic platforms and physical environments.

Future work should therefore focus on evaluating the proposed approach on real robotic platforms. In particular, experiments involving physical robot swarms would enable the practical applicability of the MCSS approach under real-world conditions to be investigated more comprehensively. In addition, extending the study to three-dimensional environment models could facilitate the analysis of more realistic search scenarios involving multilayer rubble structures and varying elevation levels. Incorporating factors such as dynamic obstacles, moving targets, sensor uncertainty, and energy consumption into the model would also allow the behavior of the approach in realistic disaster environments to be examined in greater detail.

Another important future research direction involves investigating the performance of the MCSS approach in heterogeneous robot swarm systems. Understanding how robots with different sensing capabilities, mobility characteristics, or task responsibilities can be coordinated within the same stigmergic framework represents a significant research challenge. Moreover, integrating stigmergy mechanisms with adaptive parameter update methods, learning-based movement strategies, and reinforcement learning techniques may contribute to the development of more advanced and adaptive multi-robot search systems.

Compared to the earlier thesis-based prototype framework [19], the proposed MCSS approach introduces a substantially expanded decision-making

structure and evaluation methodology. In particular, the integration of multiple environmental guidance components, comparative analysis under varying obstacle densities, and the inclusion of cost- and route-based performance metrics provide a more comprehensive assessment of stigmergy-based search behavior in complex rubble environments.

In conclusion, this study demonstrates the potential of stigmergy-based multi-robot search approaches for post-earthquake search and rescue problems within a simulation environment and shows that the proposed Multi-Component Stigmergic Search approach can provide an effective solution in complex search environments. The findings obtained in this work are expected to provide a useful foundation for the future development of autonomous multi-robot search and rescue systems.

References

[1] Liu, Y., & Nejat, G. (2013). Robotic urban search and rescue: A survey from the control perspective. *Journal of Intelligent & Robotic Systems, 72*, 147–165.

[2] Drew, D. S. (2021). Multi-agent systems for search and rescue applications. *Current Robotics Reports, 2*, 189–200.

[3] Sinha, S., Lee, S., & Singh, S. (2026). Survey on Reconnaissance Autonomous Robotic Systems for Disaster Management. *Sensors (Basel, Switzerland)*, *26*(5), 1659.

[4] Yan, Z., Jouandeau, N., & Cherif, A. A. (2013). A survey and analysis of multi-robot coordination. *International Journal of Advanced Robotic Systems, 10*(12), 399.

[5] Theraulaz, G., & Bonabeau, E. (1999). A brief history of stigmergy. *Artificial Life, 5*(2), 97–116.

[6] Karabulut, Z. (2025). *Application of a Control Algorithm Developed Based on the Ant System to the Debris Search Problem Together with a Stigmergy-Based Multi-Robot System*. M.Sc. Thesis, Bolu Abant Izzet Baysal University.

[7] Casper, J., & Murphy, R. R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B, 33*(3), 367–385.

[8] Moniruzzaman, M. D., Rassau, A., Chai, D., & Islam, S. M. S. (2022). *Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey*. Robotics and Autonomous Systems, 150, 103973.

[9] Finzi, A., & Orlandini, A. (2005). Human-robot interaction through mixed-initiative planning for rescue and search rovers. In ***AI***IA 2005: Advances in Artificial Intelligence* (pp. 483–494). Springer.

[10] Tang, H., Cao, X., Song, A., Guo, Y., & Bao, J. (2009). Human-robot collaborative teleoperation system for semi-autonomous reconnaissance robot. In *2009 International Conference on Mechatronics and Automation* (pp. 1934–1939). IEEE.

[11] Couceiro, M. S., Vargas, P. A., Rocha, R. P., & Ferreira, N. M. F. (2014). Benchmark of swarm robotics distributed techniques in a search task. *Robotics and Autonomous Systems, 62*(2), 200–213.

[12] Kumar, A. S., Manikutty, G., Bhavani, R. R., & Couceiro, M. S. (2017). Search and rescue operations using robotic Darwinian particle swarm optimization. *2017 International Conference on Advances in Computing, Communications and Informatics*, 1839–1843.

[13] Dorigo, M., Birattari, M., & Stützle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine, 1*(4), 28–39.

[14] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, 4, 1942–1948.

[15] Heylighen, F. (2016). Stigmergy as a universal coordination mechanism II: Varieties and evolution. *Cognitive Systems Research, 38*, 50–59.

[16] Payton, D., Daily, M., Estowski, R., Howard, M., & Lee, C. (2001). Pheromone robotics. *Autonomous Robots, 11*, 319–324.

[17] Fossum, F., Montanier, J. M., & Haddow, P. C. (2014). Repellent pheromones for effective swarm robot search in unknown environments. *2014 IEEE Symposium on Swarm Intelligence*, 1–8.

[18] Hamann, H., & Wörn, H. (2007). An analytical and spatial model of foraging in a swarm of robots. In *Swarm Robotics: Second International Workshop, SAB 2006* (pp. 43–55). Springer.

[19] Kegeleirs, M., & Birattari, M. (2025). Towards applied swarm robotics: current limitations and enablers. *Frontiers in Robotics and AI*, *12*, 1607978.

[20] Hunt, E. R., Jones, S. W., & Hauert, S. (2019). Testing the limits of pheromone stigmergy in high-density robot swarms. *Royal Society Open Science, 6*(11), 190225.

[21] Tang, Q., Xu, Z., Yu, F., Zhang, Z., & Zhang, J. (2019). Dynamic target searching and tracking with swarm robots based on stigmergy mechanism. *Robotics and Autonomous Systems, 120*, 103251.

[22] Chiun, J., Zhang, S., Wang, Y., Cao, Y., & Sartoretti, G. (2025, May). MARVEL: Multi-Agent Reinforcement Learning for constrained field-of-View multi-robot Exploration in Large-scale environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 11392-11398). IEEE.

[23] Li, L., Yang, B., Chen, C., Yan, Z., Sun, S., Xu, Y., ... & Zhao, L. (2025). Intelligent multi-robot exploration in non-exposed spaces: methods and challenges. *Artificial Intelligence Review*, *58*(12), 394.

[24] Wang, R., Lyu, M., & Zhang, J. (2025). A multi-robot collaborative exploration method based on deep reinforcement learning and knowledge distillation. *Mathematics*, *13*(1), 173.

Hesaplamalı Zekanın Kuramsal Temelleri: Yapay Zeka, Öğrenme Kuramı ve Büyük Veri Paradigması

Editör:
Doç. Dr. Atınç Yılmaz

ÖZGÜR YAYINLARI