ADVANCES IN COMPUTER SCIENCE:

ARTIFICIAL INTELLIGENCE, COMMUNICATION NETWORKS, AND HEALTHCARE

EDITOR:
ASST. PROF. DR. MUHAMMED NUR AVCİL

BİDGE Yayınları

BİDGE Yayınları

Advances in Computer Science: Artificial Intelligence, Communication Networks, and Healthcare

Editor: Asst. Prof. Dr. MUHAMMED NUR AVCİL

ISBN: 978-625-5474-35-3

1st Edition
Page Layout By: Gözde YÜCEL
Publication Date: 2026-11-01
BİDGE Yayınları

All rights reserved. No part of this work may be reproduced in any form or by any means, except for brief quotations for promotional purposes with proper source attribution, without the written permission of the publisher and the editor.

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

CONTENTS

Explainability in Deep Learning: Methods, Applications, and Challenges 1
YUSUF YARGI BAYDİLLİ

Ai-native 6g: Key Technologies, Applications, and Challenges .. 19
MUHAMMED NUR AVCİL

The Role of Electroencephalography in Mindfulness Research .. 36
BUSRA T. SUSAM

The Evolution of Healthcare Natural Language Processing From 2010 To 2026 56
ERKAN YAŞAR

CHAPTER 1

EXPLAINABILITY IN DEEP LEARNING: METHODS, APPLICATIONS, AND CHALLENGES

YUSUF YARGI BAYDİLLİ[1]

INTRODUCTION

Recent advances in artificial intelligence have led to the widespread use of deep learning models across a wide range of tasks. Deep learning models can directly learn complex and nonlinear patterns from data, enabling their application in various fields, including image processing, natural language processing, healthcare, finance, energy, and autonomous systems (LeCun et al., 2015). In particular, the increasing availability of large datasets and computational resources has enabled the development of deeper and more complex models.

The increasing complexity of deep learning models also introduces an important challenge. In traditional methods, the relationship between inputs and outputs can often be examined relatively easily. However, in deep learning models consisting of many layers and parameters, it is difficult to directly determine why a particular prediction is made. Therefore, deep learning models are

[1] Assist. Prof. Dr., Hakkari University, Computer Engineering, Orcid: 0000-0002-4457-2081

often described as black-box systems (Guidotti et al., 2018). A model's high predictive performance does not necessarily mean that its decision-making process is reliable. A model may produce accurate predictions by relying on biases in the data and/or features that are not directly related to the problem.

This issue is particularly important in applications where model outputs are used to make critical decisions. For example, when a model analyzes a medical image to predict a disease, it is also necessary to identify which regions of the image contribute to the prediction. Similarly, understanding which variables influence model decisions in financial risk assessment or which regions of an object a model focuses on in image classification helps provide a better understanding of the behavior of these systems. In this context, explainability is also important for researchers, as it helps identify and analyze errors during the model development process.

Numerous methods have been developed in the literature to explain deep learning models (Samek et al., 2017). Local Interpretable Model-agnostic Explanations (LIME) aims to explain individual predictions of complex models using simple, interpretable surrogate models, while SHapley Additive exPlanations (SHAP) quantifies the contribution of each feature to a model prediction (Ribeiro et al., 2016; Lundberg & Lee, 2017). Gradient-weighted Class Activation Mapping (Grad-CAM) enables the visualization of image regions that contribute to model predictions, particularly in convolutional neural networks (CNNs) (Selvaraju et al., 2017). Saliency Maps and Integrated Gradients examine the relationship between model outputs and input features using gradient-based attribution techniques (Simonyan et al., 2013; Sundararajan et al., 2017). Deep Learning Important FeaTures (DeepLIFT) evaluates the contributions of input features to model outputs by comparing neuron activations with reference activations (Shrikumar et al., 2017). Layer-wise Relevance Propagation (LRP) assigns relevance

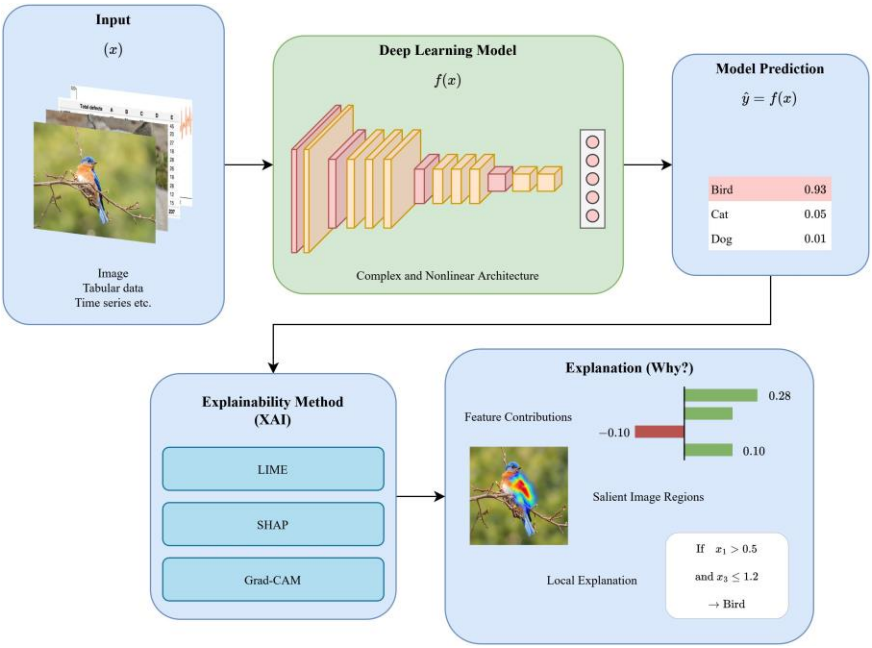
scores to input features by propagating relevance backward through the network layers (Bach et al., 2015). Other explainability methods include SmoothGrad, Occlusion Sensitivity, Anchors, and Counterfactual Explanations (Smilkov et al., 2017; Zeiler & Fergus, 2014; Ribeiro et al., 2018; Wachter et al., 2018).

Each of these methods offers a different approach to explainability. The choice of method depends on the model architecture, data type, and purpose of the explanation. Accordingly, this book chapter focuses on LIME, SHAP, and Grad-CAM due to their widespread use and different approaches to generating explanations. The chapter first introduces the concept of explainability in deep learning models and its fundamental approaches. It then examines the underlying principles, application areas, advantages, and limitations of these methods. In addition, the main challenges associated with their use are discussed to highlight the importance of explainability in understanding the decision-making processes of deep learning models.

EXPLAINABLE ARTIFICIAL INTELLIGENCE AND FUNDAMENTAL CONCEPTS

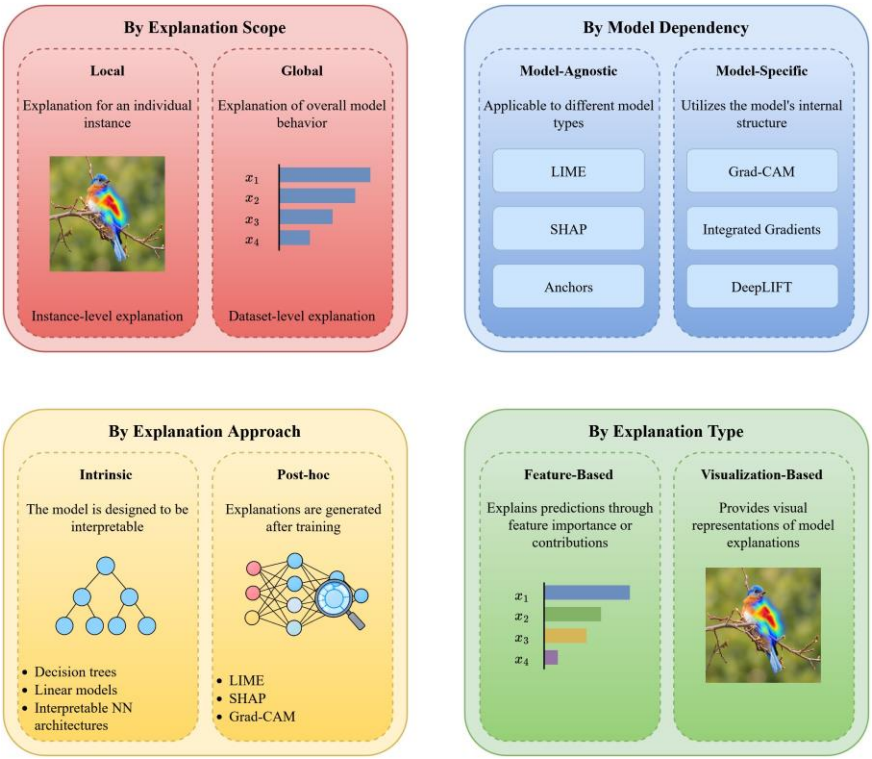
Explainable Artificial Intelligence (XAI) refers to methods and approaches that enable humans to understand and interpret the outputs of machine learning models (Barredo Arrieta et al., 2020). In deep learning models, the large number of layers and parameters, together with nonlinear operations, makes it difficult to directly understand how decisions are made. XAI methods aim to make these complex decision-making processes more understandable by identifying the inputs and features that influence model predictions (Figure 1).

Figure 1. Overview of Explainable Artificial Intelligence in Deep Learning Models.



Interpretability is another concept closely related to explainability (Lipton, 2018). Although these two terms are sometimes used interchangeably in the literature, there is a distinction between them. Interpretability refers to the ability to directly understand how a model works and how its inputs relate to its outputs. For example, linear models and decision trees can be considered interpretable because their structures can be directly examined. Explainability, on the other hand, often involves the use of additional methods to make the decisions of complex models easier to understand, particularly when these models are difficult to interpret directly. From this perspective, XAI helps explain the reasons behind the predictions made by deep learning models without requiring their internal structures to be fully simplified.

Figure 2. Classification of Explainable Artificial Intelligence Methods Based on Their Underlying Explanation Mechanisms.



Explainability methods can be classified according to different characteristics (Figure 2) (Barredo Arrieta et al., 2020). Global explainability aims to understand the overall behavior of a model across an entire dataset. This approach examines which features are most important to the model and how model outputs respond to changes in input features. Local explainability, on the other hand, focuses on explaining why a model produces a particular prediction for an individual instance. For example, local explainability can be used to understand why an image classification model assigns a specific image to a particular class or why a regression model predicts a certain value for a given instance.

Another important classification distinguishes between model-agnostic and model-specific explainability methods. Model-agnostic methods generate explanations by examining the relationship between model inputs and outputs without relying on the model's internal architecture. LIME and model-agnostic implementations of SHAP are widely used examples of this approach. Model-specific methods, on the other hand, use information from the model's internal structure, activations, or gradients to generate explanations. For example, Grad-CAM belongs to this category because it uses feature maps and gradients from the convolutional layers of CNNs.

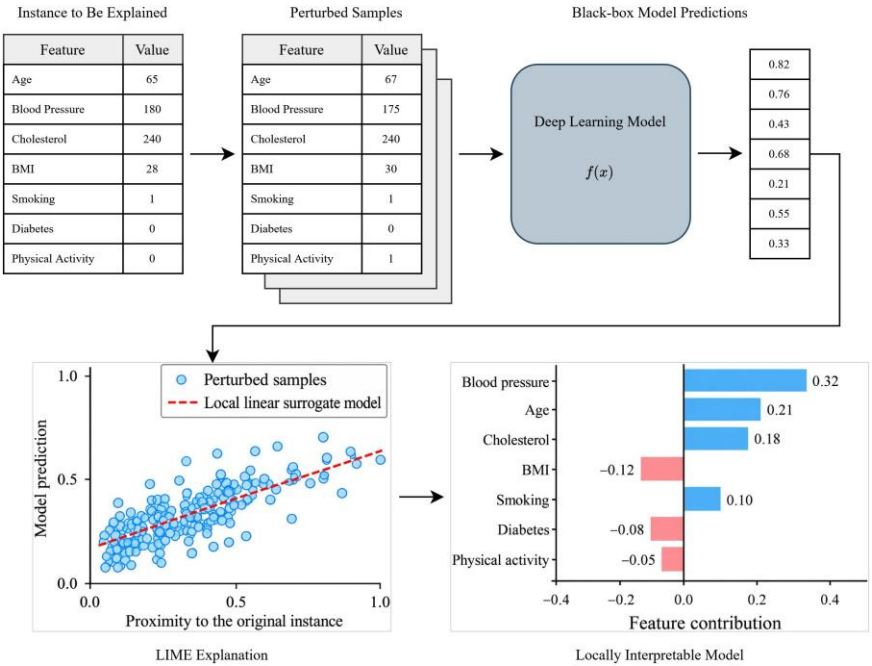
XAI methods can also be classified into intrinsically interpretable and post-hoc approaches based on how explainability is achieved. In intrinsically interpretable approaches, models are designed so that their decision-making processes can be directly understood. In contrast, post-hoc methods analyze model predictions using explanation techniques after the model has been trained. Due to the complexity of deep learning models, post-hoc methods are more commonly used to explain their predictions.

Considering these classifications together, it is clear that no single method is suitable for all models and applications. Therefore, understanding the underlying principles and limitations of XAI methods is essential for correctly interpreting the explanations they provide.

FUNDAMENTAL EXPLAINABILITY METHODS IN DEEP LEARNING

Various explainability methods have been developed to identify the features or input regions that influence the predictions of deep learning models. LIME, SHAP, and Grad-CAM are among the most widely used methods for this purpose.

Figure 3. Illustration of the LIME Framework for Explaining Individual Model Predictions.



LIME

LIME is a method that aims to locally explain individual predictions made by complex machine learning or deep learning models (Figure 3). The main idea behind LIME is to approximate the behavior of a complex model using a simpler, interpretable surrogate model trained on perturbed samples generated around the instance being explained.

Let $f(x)$ denote a deep learning model and x the instance to be explained. LIME first generates new samples by applying perturbations around x. These perturbed samples are then passed through the original model to obtain predictions. Next, the samples are weighted according to their proximity to the original instance.

Finally, a simpler and more interpretable model, such as a linear model or a decision tree, is trained using the weighted samples.

The optimization objective of LIME can be generally expressed as shown in Equation (1):

$$\xi(x) = \arg\min_{g \in G}[L(f, g, \pi_x) + \Omega(g)] \quad (1)$$

Here, f represents the complex model to be explained, g is the interpretable local surrogate model, and L measures the local fidelity loss between the surrogate model and the original model, taking into account the proximity weighting function π_x. The term $\Omega(g)$ controls the complexity of the explanation model. One of the main advantages of LIME is that it does not require access to the internal structure of the original model. Therefore, it can be applied to different machine learning and deep learning models.

SHAP

SHAP is a method that uses Shapley values from game theory to determine the contribution of each feature to model predictions (Lundberg et al., 2019). In this approach, input features are treated as players, while the model prediction is considered the outcome of their collective contributions. The aim is to calculate the contribution of each feature to the final prediction (Figure 4).

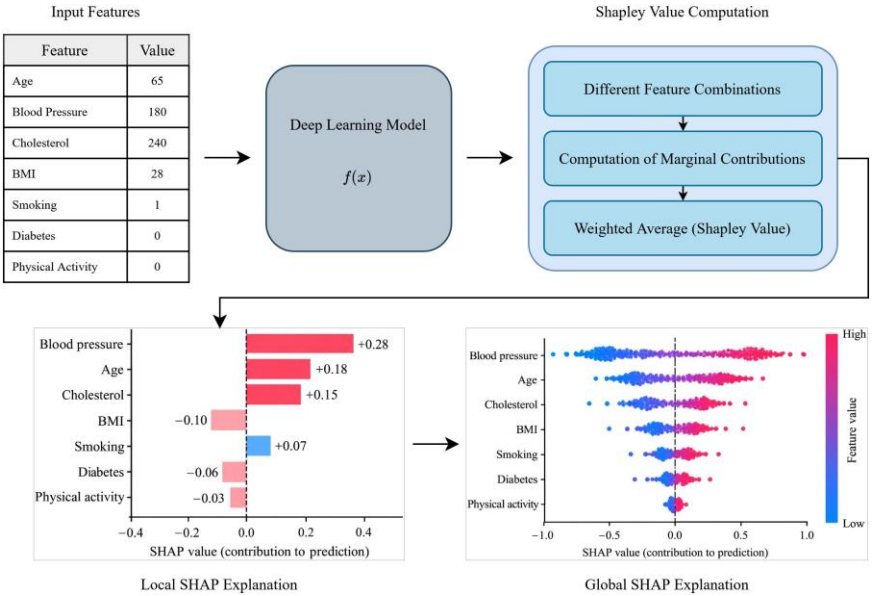
In the SHAP approach, a model prediction can generally be expressed as shown in Equation (2):

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad (2)$$

Here, ϕ_0 represents the expected model output (baseline value), while ϕ_i denotes the contribution of the i-th feature to the prediction. Positive SHAP values indicate features that increase the

model output relative to the baseline value, whereas negative values indicate features that decrease it.

Figure 4. Illustration of the SHAP Framework for Local and Global Feature Attribution.



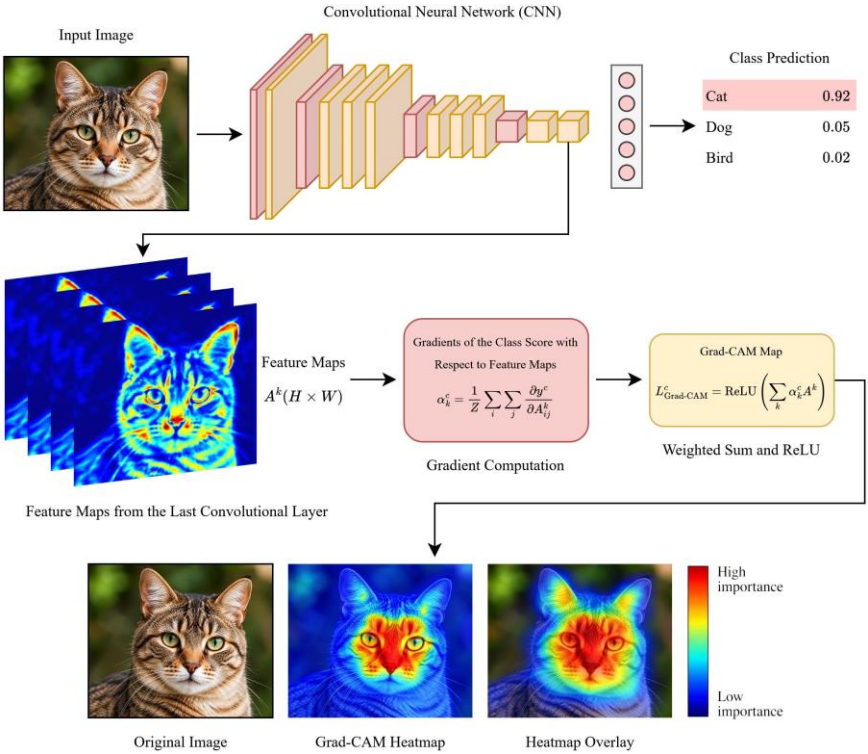
The Shapley value of a feature is calculated by taking the weighted average of the changes in model predictions when that feature is added to different combinations of other features. This approach makes it possible to examine both which features are important and how they influence a particular prediction. SHAP values can be used to explain individual predictions or aggregated across multiple samples to analyze the overall behavior of a model. Therefore, SHAP is widely used for both local and global explanations.

Grad-CAM

Grad-CAM is a method primarily used to explain image-based predictions made by convolutional neural networks (CNNs). It generates class-specific activation maps that highlight the image

regions that contribute most strongly to a particular model prediction.

Figure 5. Illustration of the Grad-CAM Framework for Visualizing Class-Discriminative Regions in Convolutional Neural Networks.



In this method, gradients of the target class score are computed with respect to the feature maps of the last convolutional layer. For a given class c, the importance weight of feature map k is calculated using Equation (3):

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

Here, y^c represents the model score for class c, A^k denotes the corresponding feature map, and Z is the number of spatial locations in the feature map. In the next step, the feature maps are combined using their importance weights to generate the Grad-CAM activation map, as shown in Equation (4):

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{4}$$

The ReLU function highlights regions that have a positive influence on the target class score (Selvaraju et al., 2017). The resulting activation map is overlaid on the original image, allowing visual examination of the regions that contribute to the model's prediction. Grad-CAM is particularly useful for evaluating model behavior in CNN-based image classification tasks. For example, Grad-CAM maps can be used to investigate whether a model focuses on the actual object or on an irrelevant background region when classifying an image.

COMPARISON AND LIMITATIONS OF EXPLAINABILITY METHODS

LIME, SHAP, and Grad-CAM use different approaches to help understand the decision-making processes of deep learning models (Table 1). While LIME and SHAP can be applied largely independently of model architecture, Grad-CAM is primarily designed for CNN-based architectures and image-based applications. The explanations produced by these methods also differ. LIME provides a local approximation of the model's behavior for a specific prediction, whereas SHAP quantitatively measures the contribution of each feature to the model output. Grad-CAM, on the other hand, visually highlights the spatial regions that contribute to the model's prediction. Therefore, the choice of explainability

method depends on the data type, model architecture, and the information expected from the explanation.

Table 1. Comparison of LIME, SHAP, and Grad-CAM methods.

Characteristic	LIME	SHAP	Grad-CAM
Underlying Approach	Local Surrogate Modeling	Shapley Value-Based Feature Attribution	Gradient-Weighted Class Activation Mapping
Explanation Scope	Local	Local and Global	Local
Model Dependence	Model-Agnostic	Implementation-Dependent	Model-Specific
Tabular Data	Applicable	Applicable	Not Applicable
Image Data	Applicable	Applicable	Highly Suitable
Explanation Output	Local Feature Contributions	SHAP Values (Feature Attributions)	Class-Discriminative Localization Heatmap
Computational Cost	Moderate	Moderate to High	Relatively Low
Main Limitation	Explanation Instability	Computational Complexity	Limited Spatial Resolution

These methods also differ in terms of explanation scope and computational cost. LIME is a local method that focuses on explaining the prediction of a specific instance. However, general trends can also be examined by aggregating explanations from multiple instances. SHAP can be used for both local and global explainability. Grad-CAM, on the other hand, is a local method that visualizes the spatial regions contributing to a particular class prediction, especially in architectures such as CNNs. In terms of computational cost, LIME can be computationally expensive because it requires model predictions for a large number of perturbed samples. The computational cost of SHAP depends on the algorithm used. In particular, exact computation of Shapley values can be computationally demanding for high-dimensional data, complex models, and model-agnostic applications. In contrast, Grad-CAM generally requires only one forward pass and one backward pass for

the target class, making it less computationally expensive than many LIME and SHAP implementations (Lundberg & Lee, 2017; Ribeiro et al., 2016; Selvaraju et al., 2017).

Although explainability methods provide valuable insights into model behavior, their results should be carefully evaluated. One of the main limitations of these methods is that the explanations may not always accurately reflect the model's actual decision-making process. In particular, post-hoc methods generate explanations after predictions have been made and provide an interpretation of the model's behavior rather than a direct representation of its internal decision-making process. Therefore, the explanations produced by these methods should not be interpreted as direct evidence of causal relationships.

Another important limitation is the stability of explanations. Small changes in input data may lead to significant differences in the explanations produced by some methods. For example, in LIME, changes in the perturbations generated around the instance being explained may alter the features highlighted in the local explanation. Similarly, SHAP results can be influenced by the computational method and the reference values used. Producing substantially different explanations for similar inputs can reduce the reliability of explainability methods. In visual explanation methods, interpreting the results presents an additional challenge. Although Grad-CAM heatmaps allow visual examination of the regions that contribute to model predictions, a highlighted region should not be considered the sole reason for the model's decision. Furthermore, the limited spatial resolution of activation maps makes it difficult to identify fine details that influence the model's predictions.

Another challenge in evaluating explainability methods is the lack of a single standardized evaluation metric that can be applied to all problems. How easily an explanation can be understood by users, how accurately it represents model behavior, and how consistently it

produces results across different instances are distinct evaluation aspects. Therefore, obtaining explanations that are visually convincing and easy to interpret is not sufficient to demonstrate their reliability.

In conclusion, methods such as LIME, SHAP, and Grad-CAM are useful tools for understanding the decisions made by deep learning models. However, their outputs should not be considered definitive or complete explanations of model decisions. The reliability and stability of the explanations, as well as the assumptions underlying the methods, should be carefully considered. Furthermore, whenever possible, results obtained from different explainability approaches should also be evaluated.

CONCLUSION

Deep learning methods have been successfully applied in many fields due to their ability to learn meaningful patterns from complex data. However, the increasing complexity of model architectures makes it difficult to understand the decision-making processes underlying their predictions. In applications such as healthcare, finance, energy, and autonomous systems, where model outputs are particularly important, it is essential not only to achieve high predictive performance but also to understand the information on which model decisions are based. This need has made explainable artificial intelligence approaches an important component of deep learning research.

In this book chapter, the fundamental concepts of explainability in deep learning models have been discussed, and three widely used methods, LIME, SHAP, and Grad-CAM, have been examined. Although these methods share the same general objective, they use different explanation mechanisms. Therefore, the explanations produced by these methods may vary depending on the method used, its parameters, and the characteristics of the data.

Furthermore, generating explanations for a model does not necessarily mean that the model is fully transparent or reliable. For this reason, XAI outputs should be considered as complementary information that helps analyze and understand model behavior.

As deep learning systems are increasingly used in a wider range of critical applications, the importance of explainability is expected to grow further. In this regard, it is important not only to develop more understandable explanations but also to establish standardized approaches for evaluating their accuracy, stability, and reliability. Considering both predictive performance and explainability together will contribute to the development of more transparent and reliable artificial intelligence systems that users can better understand and evaluate.

REFERENCES

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

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

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable AI for trees: From local explanations to global understanding. *arXiv.1905.04610*. https://doi.org/10.48550/arXiv.1905.04610

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural*

Information Processing Systems 30, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Thirty-Second AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11491

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv:1708.08296*. https://arxiv.org/abs/1708.08296v1

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. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *34th International Conference on Machine Learning*, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*. https://doi.org/10.48550/arxiv.1312.6034

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *arXiv.1706.03825*. https://doi.org/10.48550/arXiv.1706.03825

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning*, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *arXiv.1711.00399*. https://doi.org/10.48550/arXiv.1711.00399

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *ECCV: European Conference on Computer Vision 2014, Lecture Notes in Computer Science*, *8689*, 818–833.

CHAPTER 2

AI-NATIVE 6G: KEY TECHNOLOGIES, APPLICATIONS, AND CHALLENGES

MUHAMMED NUR AVCİL[1]

Introduction

In recent years, mobile communication technologies have evolved from systems that were primarily designed to facilitate communication between people and have become an integral part of wider digital ecosystems, where people, machines, sensors, vehicles and physical environments are constantly interacting with each other. While the first-generation mobile communication systems mainly focused on voice communication, data communication, mobile Internet, the Internet of Things (IoT), and high-speed mobile services became more prominent with 2G and subsequent generations. Now, communication technologies such as 5G and 5G-Advanced support a wide range of requirements, including high data rates, low latency, high device connectivity, and reliable communication (Alliance & Germany, 2024).

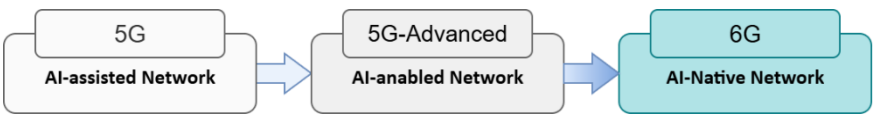
Nowadays, the rapid increase in the number of connected devices, the increasing variety of data traffic, the use of artificial

[1] Asst. Prof. Dr., Hakkari University, Faculty of Engineering, Department of Computer Engineering, Hakkari, Türkiye, Orcid: 0000-0001-8220-5576

intelligence-based applications, the growth of autonomous systems, and the increasing complexity of communication networks have created a need to evolve/develop existing network architectures to adapt to these evolving requirements. This has led to a demand for next-generation network approaches that can not only provide higher capacity, but also adapt to varying traffic conditions, utilize resources more effectively and efficiently, and meet different service requirements (Klaus & Berndt, 2018).

5G mobile communication technology, together with artificial intelligence and machine learning, is used to solve specific problems such as traffic prediction, resource allocation, energy optimization, mobility management, and network performance analysis. In 5G, the existing network architecture is largely maintained, and artificial intelligence solutions are added to the application layer of this architecture. Therefore, these methods proposed as solutions are generally considered "AI-assisted" or "AI-enabled" network approaches. In 6G mobile communication technology, however, a more advanced approach, known as the "AI-Native" network concept, has emerged (Li, 2024). As shown in Figure 1, the transformation in mobile network communications is illustrated.

Figure 1. Transformation from 5G to AI-Native 6G



With 6G, communication, computing, sensing, and artificial intelligence are intended to be considered in a more integrated structure. Among the 6G usage scenarios defined by the International Telecommunication Union (ITU) within the IMT-2030 framework

(Alliance & Germany, 2024), there is also an “AI and Communication” approach that refers to the integration of artificial intelligence and communication. This approach shows that artificial intelligence is aimed to become one of the fundamental elements of the network architecture in 6G, rather than being only a supporting technology that improves specific network functions.

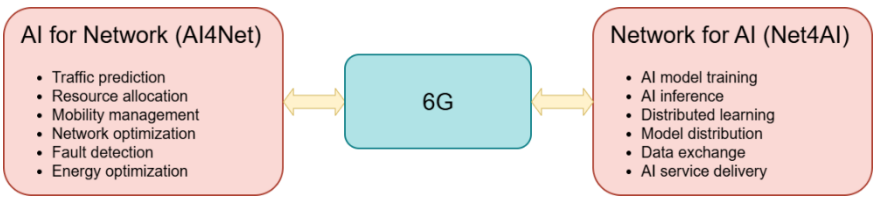
In the AI-Native approach, artificial intelligence is considered a fundamental architectural component that is taken into account at different levels, from network design to operation and from resource management to service delivery, rather than as an independent function added to the network later. The aim of this approach is not only to use artificial intelligence in the network, but also to design the network architecture by considering the requirements and capabilities of artificial intelligence (Wu et al., 2022).

The studies conducted by 3GPP consider the transformation from two perspectives. The first approach is called “AI for Network (AI4Net)”, which refers to the application of artificial intelligence to improve network performance and network management. The second approach is called “Network for AI (Net4AI)”, which refers to the communication network providing the data, connectivity, computing and service infrastructure for the requirements of artificial intelligence applications. As stated in the technical evaluations published by 3GPP (3GPP 38.745, 2026), these two approaches will be among the fundamental principles in the design of 6G networks.

The AI4Net method allows the use of artificial intelligence to improve decision-making mechanisms in the network. For example, a base station can predict its future traffic load, can anticipate the mobility of mobile users, or can dynamically allocate radio resources according to the current and future network conditions. In the Net4AI approach, the network can provide the communication and

computing infrastructure required for operations such as training distributed artificial intelligence models, transferring large datasets, or deploying artificial intelligence models on the endpoint devices and edge servers (Yuan et al., 2025). Figure 2 shows some examples use cases for AI4Net and Net4AI, as well as the relationship between AI4Net and Net4AI.

Figure 2. Relationship Between AI4Net and Net4AI



Therefore, the relationship between 6G and artificial intelligence is not one-directional. While the application of artificial intelligence makes 6G networks more intelligent, 6G networks can provide the artificial intelligence applications with lower latency, higher reliability and wider range access. This mutual relationship forms of the AI-Native 6G concept (Alhammadi et al., 2024).

Another important issue in the development of 6G is that networks are becoming more and more heterogeneous. Terrestrial base stations, small cells, satellite systems, end devices, vehicles, robots, and edge servers have become parts of the same communication ecosystem. In such a dynamic environment, with many different components, network management based on fixed rules may not be sufficient in all cases. Artificial intelligence can allow the network to continuously collect data from its environment, analyze the current state, predict future conditions and automatically make appropriate decisions (Alhashimi et al., 2023).

The main goal of AI-Native 6G is not simply to “use more artificial intelligence.” The main objective is to transform the communication infrastructure into a system that can perceive its network environment, learn, make decisions, dynamically manage its resources, and adapt to changing conditions.

AI-Based Functions and Technologies in AI-Native 6G Networks

The purpose of developing the AI-Native 6G architecture is to enable artificial intelligence to perform various functions within the network. 6G networks are expected to support a very large number of devices, varying traffic patterns, heterogeneous communication technologies, and different service requirements simultaneously. In wireless networks, resources such as spectrum, radio resource blocks, processing power, energy, and network capacity are limited. Users' resource demands also vary over time. Existing methods generally allocate these resources among users using predefined rules or optimization algorithms. On the other hand, AI-based methods can use historical data request of user demands and real-time network data situation to predict future resource requirements and make resource allocation decisions based on these predictions (Tong & Li, 2022). For example, in the urban area, vehicle density may increase during the morning and evening rush hours. A relationship can be established between the increase in traffic density and vehicle resource demand. Machine learning algorithms can use the historical data of network vehicle density and resource allocation to predict the workload that may occur in specific areas at specific times. According to the predictions, the network can plan its limited resources, including channel resource capacity and edge computing capacity (Kato et al., 2020).

Similarly, artificial intelligence can help in handover management between base stations. Current handover decisions are made using parameters like signal strength, cell load ratio and set threshold values. However, the future movements of mobile users can be predicted to make the decision of handover earlier and more accurately. Beam management is another area where artificial intelligence can be applied. Directional antennas and beamforming methods are becoming increasingly important for communication systems that operate at high frequencies. The most suitable beam for future communication can be predicted from the current location, direction of movement, speed of the user and the previous channel conditions. This can improve connection continuity and reduce redundant beam-scanning operation (Yang et al., 2020).

Network traffic prediction is also one of the important goals of AI-Native 6G. Methods such as time-series models, deep learning, and graph-based models can be used to predict data traffic by utilizing data transfer information between different base stations and users. In particular, learning methods such as graph neural networks (GNNs) are suitable for 6G networks due to their dynamic and heterogeneous structure. Recent studies indicate that GNN-based approaches can be used in 6G networks for applications such as dynamic topology management, network traffic prediction, edge computing, and network optimization (Zeb et al., 2021; Hamzaoğlu et al., 2025; SUSAM, 2025).

Federated learning is another important machine learning approach for 6G-AI integration. In centralized learning, user data is sent to a central server, whereas in federated learning, model training tasks are performed closer to the data sources, reducing the need to send raw data directly to the central system. Federated learning is particularly useful in applications with high privacy requirements. However, differences in processing capacity, data volume, and

connection quality among distributed devices create new research opportunities for federated learning (Yang et al., 2022).

Artificial intelligence can also be used for network automation, which is another important area in AI-Native 6G networks. The aim is for the network to be constantly self-monitoring, self-diagnosing, self-managing its resources and self-optimizing its performance, with no need for human control. This approach is related to Zero-touch Network and Service Management (ZSM), which aims to automate network and service management with minimal human intervention. It enables the network to automate the processes of monitoring, analysis, decision-making, and action with minimal human involvement (Majumdar et al., 2026).

In this context, artificial intelligence models are intended to be used not only for making predictions but also for making decisions and controlling network behavior. Methods such as reinforcement learning can help the network develop appropriate decisions by observing the outcomes of different actions. However, for network functions where reliability is critical, the validation, explainability, and safe implementation of AI-based decisions remain important issues that require further investigation.

Thus, the application of AI in AI-Native 6G networks is not limited to a single network function. Many different network functions such as resource allocation, mobility management, traffic forecasting, energy optimization, etc., will be possible with AI-Native 6G through an artificial intelligence infrastructure.

AI-Native 6G Applications, Security, and Research Challenges

AI-Native 6G offers new opportunities to support mobile communication applications across various sectors. Intelligent transportation systems, autonomous vehicles, robotics, smart cities,

industrial automation, healthcare services, large-scale IoT systems, and digital twins are among the main application areas (Cui et al., 2025).

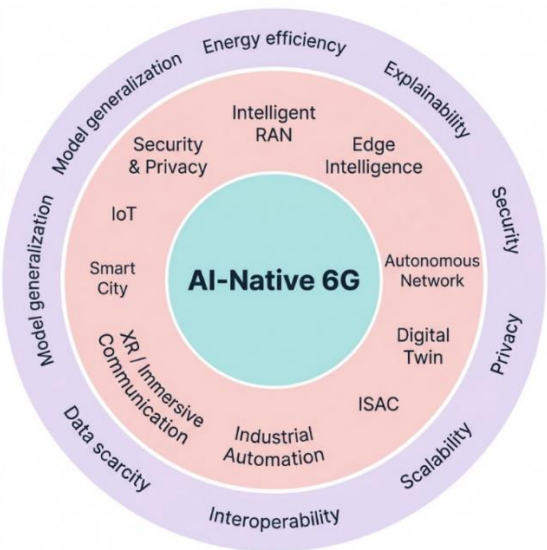
V2X communication is a particularly important application area for AI-Native 6G. Connected and autonomous vehicles are expected to continuously exchange data not only with other vehicles but also with road infrastructure, traffic lights, edge servers, and mobile networks. Information such as vehicle location, speed, direction, traffic density, sensor data, and road conditions can be used as input for artificial intelligence models. With the AI-Native 6G approach, the communication requirements of vehicles can be evaluated together with the limited network resources. For example, an increase in data requests can be predicted as the number of vehicles at a busy intersection increase, allowing the resources of the relevant access points to be adjusted in advance to accommodate the expected increase in demand. Similarly, predicting the future positions of vehicles can help improve handover decisions and edge server selection. The AI-Native 6G architecture can enable communication resources to be managed not only according to current demand but also based on future demand and user mobility information (Kandali & Nouh, 2026).

Another one of the key application area of AI-Native 6G is smart cities. Smart cities are based on a massive number of devices that are constantly producing data e.g. traffic light management, energy systems, security cameras, environmental sensors, public transportation systems, etc. The data can be analyzed through artificial intelligence that allows a more dynamic and efficient management of the city infrastructure. Robots and machines in industrial applications need reliable, low-latency communication. Artificial intelligence can be used for analysis of the operating conditions of robots in the production environment, prediction of maintenance needs and optimization of production processes. The

main goal here is for the 6G network to be not just an infrastructure for data transmission, but a distributed platform that enables the operation of AI services (Ogenyi et al., 2025).

Digital twin technologies are also an important part of the AI-driven transformation. Creating a digital copy of a physical system and continuously updating it with data from the real system can be useful for network management and the optimization of physical systems. For example, a digital twin of a transportation system can be used to simulate traffic density, vehicle movements, and data communication traffic. It can also be used to evaluate the potential outcomes of solutions that cannot be tested directly in the physical environment. By analyzing the outcomes of different decisions in this digital environment, artificial intelligence can help develop suitable solutions that can be applied to the real system (Lin et al., 2023).

Figure 3. Applications and Research Challenges in AI-Native 6G Networks



AI-Native 6G networks will process large amounts of user data and network traffic data. Therefore, protecting the privacy of this data is essential. In particular, information about location, mobility, health, vehicles and user behavior may contain personal or sensitive information and it needs to be protected. AI models themselves are also vulnerable to attacks. For example, a model can be tricked into making the wrong decisions by manipulation of training data. Adversarial attacks cause the model to misclassify an input that it would normally classify correctly. Therefore, AI-Native 6G requires not only secure communication but also the protection of AI models and their training processes (Ogenyi et al., 2025).

Table 1. Key Research Challenges and Possible Approaches in AI-Native 6G

Challenge	**Importance**	**Possible Approach**
Data privacy	User data	Federated Learning
Latency	Real-time applications	Edge AI
Energy Consumption	Cost of AI models	Model Compression
Security	Adversarial attacks	Robust AI
Explainability	Critical decisions	Explainable AI
Scalability	Large number of devices	Distributed AI
Updating	Dynamic network environment	Continual Learning

Another important challenge in AI-integrated networks is explainability. When an AI model used within the network makes a critical decision regarding handover, resource allocation, or traffic routing, network administrators may need to understand the reasons behind that decision. For this reason, explainable artificial intelligence and trustworthy AI are among the important research areas for AI-Native 6G. Training and running AI models require significant computational and energy resources. Resources are

particularly limited on end devices and edge servers. Therefore, a balance must be maintained between the computational cost, latency, and energy consumption associated with training and running AI models (Alhammadi et al., 2024; Cui et al., 2025). In Table 1 list some research challenges and possible approaches for AI-Native 6G networks.

In addition, network components developed by different companies need to operate within a common AI infrastructure. To make 6G a global standard, AI models, data formats, interfaces, and network functions must be able to work together within a standard. Appropriate standards are therefore needed to ensure interoperability among network components. Standardization activities on AI for 6G are already ongoing. For example, 3GPP has initiated a Release 20 (3GPP 38.745, 2026) study on the protocol for AI in 6G, and the related technical report remains in draft status. Release 20 also includes work on AI/ML management and AI/ML support in the core network. Therefore, AI-Native 6G is not yet a fully developed technology; rather, it is an evolving field in which standardization, prototyping, experimental studies, and academic research are progressing together.

Conclusion and Future Perspectives

6G aims to go beyond improvements in network capacity and data rates by making mobile communication networks more intelligent, more adaptable, distributed, and autonomous. In this process, artificial intelligence is expected to become a fundamental part of the network architecture rather than serving only as a supporting technology.

Although artificial intelligence and machine learning have started to be used to solve specific network problems in 5G and 5G-Advanced networks, the AI-Native 6G approach represents a more

comprehensive transformation. In this approach, the network not only uses artificial intelligence but also provides the communication and computing infrastructure required for AI applications to operate. Therefore, considering the AI4Net and Net4AI approaches together is important for understanding future 6G network architectures.

In the AI-Native 6G architecture, artificial intelligence can be utilized in multiple domains such as user devices, radio access network, core network, edge computing and cloud infrastructure. This distributed approach can enable: More efficient use of network resources, Prediction of traffic patterns and user behavior, better mobility management, less energy consumption, more automation of network operations. In particular, edge intelligence is emerging as an important part of AI-native 6G. Running AI models close to data sources is expected to help reduce latency and communication overhead while improving data privacy in some applications by reducing the need to transfer data to centralized systems. However, new research challenges are expected to arise in areas such as the management of distributed AI systems, model updates, resource sharing, and coordination among different edge nodes.

One of the important application areas of AI-Native 6G is V2X and autonomous driving systems. Artificial intelligence models can enable more dynamic management of network resources by analyzing vehicle mobility, data demands, and environmental conditions. Considering this, the AI-Native 6G approach may help develop more effective solutions to problems such as vehicle density, data offloading, edge node selection, resource allocation, handover, and traffic prediction.

With AI-Native 6G, the aim is to move beyond network systems that respond only to user data requests and develop systems that can predict future user conditions and plan the required resources in advance. For this reason, different methods, including machine learning, deep learning, reinforcement learning, federated

learning, graph neural networks, and generative artificial intelligence, may be used together in the near future. Current research also focuses on adapting these methods to the dynamic and heterogeneous nature of 6G networks.

The aim of AI-Native 6G is to move beyond network systems that only respond to user data requests and to develop systems that can predict future user conditions and plan the required resources in advance. For this reason, different methods such as machine learning, deep learning, reinforcement learning, federated learning, graph neural networks, and generative artificial intelligence may be used together in the near future. Research is also underway to adapt these methods to the dynamic and heterogeneous nature of 6G networks. However, simply developing highly accurate AI models is not enough to make AI-Native 6G a success. These models need to operate with low latency and high energy efficiency, be able to make reliable decisions, be protected against attacks, and work together with other network components.

To summarize, AI-Native 6G is an approach in where communication and artificial intelligence become intertwined and supportive of each other parts of the network architecture. Future 6G networks are expected to be more than just infrastructures providing faster connections. They are designed to sense the environment, learn from data, predict future conditions, dynamically manage resources and adapt to different applications.

In this sense, AI-Native 6G is a broad research area that combines communication, artificial intelligence, edge computing, sensing, and automation technologies. As these technologies mature in and beyond 2030, networks are expected to become more autonomous and context-aware. In this process, artificial intelligence is expected to evolve from a service provided over the network into one of the fundamental elements shaping how the network operates.

References

3GPP 38.745, R.-2. (2026). *Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN Phase 3.* 3GPP.

Alhammadi, A., Shayea, I., El-Saleh, A. A., Azmi, M. H., Ismail, Z. H., Kouhalvandi, L., & Saad, S. A. (2024). Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges. *International Journal of Intelligent Systems*, 8845070. https://doi.org/10.1155/2024/8845070

Alhashimi, H. F., Hindia, M. N., Dimyati, K., Hanafi, E. B., Safie, N., Qamar, F., Azrin, K., & Nguyen, Q. N. (2023). A survey on resource management for 6G heterogeneous networks: Current research, future trends, and challenges. *Electronics, 12*(3), 647. https://doi.org/10.3390/electronics12030647

Alliance, N., & Germany, D. (2024). *ITU-R framework for IMT-2030: Review and future direction.* NGMN Alliance.

Cui, Q., You, X., Wei, N., Nan, G., Zhang, X., Zhang, J., Lyu, X., Ai, M., Tao, X., Feng, Z., & others. (2025). Overview of AI and communication for 6G network: Fundamentals, challenges, and future research opportunities. *Science China Information Sciences, 68*(7), 171301. https://doi.org/10.1007/s11432-024-4337-1

Hamzaoğlu, A., Erduman, A., & Kırçay, A. (2025). Deep Learning-Based Rooftop PV Detection and Techno Economic Feasibility for Sustainable Urban Energy Planning. *Sustainability, 17*(15), 6853. https://doi.org/10.3390/su17156853

Kandali, K., & Nouh, S. (2026). AI-Native V2X and Internet of Vehicles Systems: Architectures, Learning Paradigms, and

Open Challenges. *IEEE Open Journal of Intelligent Transportation Systems, 7*, 1775 - 1788. https://doi.org/10.1109/OJITS.2026.3706124

Kato, N., Mao, B., Tang, F., Kawamoto, Y., & Liu, J. (2020). Ten Challenges in Advancing Machine Learning Technologies toward 6G. *IEEE Wireless Communications, 27*(3), 96-103. https://doi.org/10.1109/MWC.001.1900476

Klaus, D., & Berndt, H. (2018). 6G vision and requirements: Is there any need for beyond 5G? *IEEE vehicular technology magazine, 13*(3), 72-80. https://doi.org/10.1109/MVT.2018.2848498

Li, L. (2024). A Survey on Intelligence-Endogenous Network: Architecture and Technologies for Future 6G. *Intelligent and Converged Networks, 5*(1), 53-67. https://doi.org/10.23919/ICN.2024.0005

Lin, X., Kundu, L., Dick, C., Obiodu, E., Mostak, T., & Flaxman, M. (2023). 6G digital twin networks: From theory to practice. *IEEE Communications Magazine, 61*(11), 72-78. https://doi.org/10.1109/MCOM.001.2200830

Majumdar, S., Wei, Q., Schwarzmann, S., Trivisonno, R., & Carle, G. (2026). Toward AI-Native 6G Systems: Standards Enablers for 6G Network Automation. *IEEE Communications Standards Magazine, 10*(1), 145-153. https://doi.org/10.1109/MCOMSTD.2025.3585376

Ogenyi, F. C., Ugwu, C. N., & Ugwu, O. P.-C. (2025). A comprehensive review of AI-native 6G: integrating semantic communications, reconfigurable intelligent surfaces, and edge intelligence for next-generation connectivity. *Frontiers in Communications and Networks, 6*, 1655410. https://doi.org/10.3389/frcmn.2025.1655410

SUSAM, B. T. (2025). ENSEMBLE LEARNING MODELS FOR ANALYZING FACTORS INFLUENCING MATHEMATICS ACHIEVEMENT IN PISA 2022 TÜRKIYE. In *COMPUTER ENGINEERING* (p. 7). Gece Kitaplığı.

Tong, W., & Li, G. Y. (2022). Nine Challenges in Artificial Intelligence and Wireless Communications for 6G. *IEEE Wireless Communications, 29*(4), 140-145. https://doi.org/10.1109/MWC.006.2100543

Wu, W., Zhou, C., Li, M., Wu, H., Zhou, H., Zhang, N., Shen, X. S., & Zhuang, W. (2022). AI-Native Network Slicing for 6G Networks. *IEEE Wireless Communications, 29*(1), 96-103. https://doi.org/10.1109/MWC.001.2100338

Yang, H., Alphones, A., Xiong, Z., Niyato, D., Zhao, J., & Wu, K. (2020). Artificial-Intelligence-Enabled Intelligent 6G Networks. *IEEE Network, 34*(6), 272-280. https://doi.org/10.1109/MNET.011.2000195

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

Yuan, Y., Sun, B., Zeng, J., Wu, Q., Liu, J., Jiang, D., & Qin, F. (2025). 6G Network Architecture: QoS Paradigms and Data Lifecycle Management for Next-Generation Networks. *IEEE Communications Magazine, 63*(8), 16-22. https://doi.org/10.1109/MCOM.001.2400564

Zeb, S., Rathore, M. A., Mahmood, A., Hassan, S. A., Kim, J., & Gidlund, M. H. (2021). Edge Intelligence in Softwarized 6G: Deep Learning-enabled Network Traffic Predictions. (pp. 1-

6). Spain: IEEE. https://doi.org/10.1109/GCWkshps52748.2021.9682131

CHAPTER 3

THE ROLE OF ELECTROENCEPHALOGRAPHY IN MINDFULNESS RESEARCH

BUSRA T. SUSAM[1]

Introduction

Mindfulness involves intentional awareness of present-moment experience with a nonjudgmental orientation and can be cultivated through formal meditation and everyday practice. A primary goal in mindfulness research is to relate subjective experiences to measurable behavioral and neural processes. Self-report questionnaires assess participants' reported experiences, while experimental tasks assess behavioral performance under controlled conditions. Neurophysiological methods complement these approaches by characterizing brain activity associated with mindfulness practice and training. These complementary measures allow researchers to examine whether differences in brain activity are accompanied by differences in task performance and participants' reported experiences.

Electroencephalography (EEG) contributes to this investigation by measuring changes in electrical potentials at the

[1] Assistant Professor, Hakkari University, Faculty of Engineering, Department of Computer Engineering, Orcid:0000-0001-8250-2044

scalp that reflect the coordinated activity of large populations of neurons. Its millisecond temporal resolution allows researchers to examine how neural activity unfolds over time. In mindfulness research, EEG can be recorded during meditation to characterize ongoing patterns of activity or during experimental tasks to examine responses to specific stimuli. These recordings provide complementary information about activity associated with the practice itself and the timing of cognitive or emotional processing assessed outside meditation.

This chapter presents a narrative overview of EEG approaches in mindfulness research. It introduces the conceptual foundations of mindfulness and meditation, examines practice types and practitioner experience, and briefly considers complementary measurement methods. It then discusses oscillatory activity followed by methodological challenges and priorities for future research. The chapter aims to clarify how EEG contributes to understanding mindfulness and how experimental design shapes the interpretation of neural findings.

Conceptual Foundations of Mindfulness and Meditation

Meditation encompasses structured practices with different instructions, attentional demands, and purposes. Some focus attention on an object, such as the breath, while others encourage awareness of changing sensations, thoughts, and emotions. Mindfulness describes intentional, nonjudgmental awareness of present-moment experience, cultivated through meditation or everyday activities such as walking, eating, and conversation(Kabat-Zinn, 2003)

In psychological and neuroscientific research, mindfulness provides a framework for examining how individuals process and respond to experience. (Bishop et al., 2004) dentified two components: regulating attention toward present-moment

experience and approaching it with curiosity, openness, and acceptance. Mindfulness therefore involves both awareness and the attitude brought to it, distinguishing it from concentration or simply noticing a stimulus. For example, noticing shoulder tension reflects attention, while observing the sensation and accompanying thoughts or emotions with acceptance reflects a mindful orientation.

The relationship between meditation and mindfulness depends on how a practice is performed. During mindful walking, individuals attend to movement and contact with the ground, notice distractions, and gently return attention to the experience. Similarly, formal meditation involves more than sitting or closing the eyes; instructions and engagement help define the mental processes involved (Lutz et al., 2008). Mindfulness-based programs should therefore be described in terms of its specific exercises, teaching approach, and opportunities for practice (Crane et al., 2017; Loucks et al., 2022; Tang & Posner, 2013). In EEG research, these details clarify participants' activities during recording and support meaningful comparisons across studies (Lee et al., 2018; Lomas et al., 2015).

Assessment Methods in Mindfulness Research

Several measurement approaches are employed in mindfulness research to investigate reported experience, behavioral performance, and associated brain activity. These approaches address different aspects of the research question. Questionnaires assess how individuals describe their awareness and experiences. Behavioral tasks examine performance under defined experimental conditions. Neuroimaging methods investigate neural correlates of mindfulness practice and results. Therefore, findings should be interpreted according to what each method measures, rather than treated as interchangeable indicators of mindfulness.

Self-report questionnaires provide a structured means of assessing how individuals describe their attention, awareness, and relationship with their experiences. The Mindful Attention Awareness Scale measures how frequently individuals report being attentive to and aware of present-moment experiences (Brown & Ryan, 2003). It emphasizes variation in attention and awareness over time, rather than qualities such as empathy, acceptance or gratitude. The Five Facet Mindfulness Questionnaire extends assessment beyond attention and awareness to include how individuals describe and relate to their experiences through five facets such that observing, describing, acting with awareness, nonjudging of inner experience, and nonreactivity to inner experience (Baer et al., 2006). Although, both questionnaires provide information about participants' attention, awareness, and responses to experience, their interpretation depends on how participants understand the items and describe their experiences.

Whereas questionnaires assess participants' descriptions of their experiences, behavioral tasks investigate specific cognitive processes that may be relevant to mindfulness, including attentional control and the allocation of processing resources. Outcomes such as response accuracy and reaction time provide information about performance under defined experimental conditions. These tasks help translate a proposed psychological process into an observable outcome, although they do not measure mindfulness as a complete construct. (Moore et al., 2012) used the Stroop Word–Color Task, in which participants identified a word's ink color while ignoring its meaning, to examine whether mindfulness training influenced the ability to manage conflicting information. (Slagter et al., 2007) used an attentional-blink task to investigate the detection of two targets presented close together in time. The second target is frequently missed when it appears shortly after the first. Following a three-month intensive meditation training, participants showed a greater

reduction in the attentional-blink deficit than a novice comparison group. Behavioral tasks therefore provide performance-based measures for examining whether mindfulness training is associated with changes in specific cognitive abilities.

Brain imaging extends this assessment from cognitive performance to the neural processes and anatomical characteristics associated with mindfulness. Functional magnetic resonance imaging (fMRI) measures blood-oxygenation changes associated with neural activity, allowing researchers to investigate the involvement of brain regions and networks during tasks, meditations or resting conditions. (Brewer et al., 2011; Farb et al., 2007; Logothetis et al., 2001) For instance, (Farb et al., 2007) utilized fMRI to investigate narrative and experiential forms of self-focused attention, finding that participants with mindfulness training showed more pronounced reductions in medial prefrontal activity and greater engagement of regions involved in bodily awareness when attending to present moment experiences rather than reflecting on personal traits. Structural MRI examines anatomical characteristics including cortical thickness and regional gray matter volume or concentration, allowing researchers to investigate whether meditation experience is associated with differences in brain anatomy and whether structural changes occur following meditation exercise (Hölzel et al., 2011; Kral et al., 2022; Lazar et al., 2005) . For instance, (Lazar et al., 2005) used structural MRI and found greater cortical thickness in the prefrontal cortex and right anterior insula in experienced meditators compared with matched control. (Hölzel et al., 2011) utilized structural MRI before and after an eight-week mindfulness-based stress reduction program and reported increases in gray matter concentration in the left hippocampus, posterior cingulate cortex, temporoparietal junction, and cerebellum compared with a waiting-list control group.

The approaches described above provide different perspectives on mindfulness, while EEG allows closer examination of how electrical brain activity unfolds over time. Recordings during meditation, rest, and cognitive tasks support the study of brain oscillations and responses to specific events, complementing information about reported experience, behavioral performance and brain anatomy. EEG findings should be interpreted in relation to practice instructions, recording conditions, and the research question, because they do not directly measure mindfulness as a complete construct. The following section examines the main EEG measures used in mindfulness research and their associated findings.

EEG Measures and Findings in Mindfulness Research

EEG records electrical potential differences at the scalp, allowing researchers to examine brain activity over time. In mindfulness research, EEG recordings are obtained during meditation, rest, and cognitive tasks. Whereas recordings during meditation characterize activity accompanying the practice, recordings during rest or tasks can examine differences outside the meditation session. For example, (Cahn et al., 2010) compared EEG activity during meditation and a mind-wandering control condition in the same experienced meditators, examining differences between mental states. By comparison, (Nyhus et al., 2019) recorded EEG during a memory retrieval task before and following four weeks of mindfulness training, with a waiting-list control group. These examples illustrate how EEG study designs can examine both activities accompanying meditation and changes in task-related activity over a training period.

EEG measures describe several properties of the recorded signal. Spectral analyses examine how power is distributed across frequencies; event-related potentials characterize responses to specific events; connectivity analyses assess statistical relationships

between signals; and complexity measures describe aspects of signal organization. These approaches provide complementary information, but their findings should be interpreted in relation to the practice, comparison condition, participant experience, and analysis method.

Oscillatory Activity

Spectral analysis characterizes how EEG signal power is distributed across frequencies by estimating the power spectral density (PSD). Welch's method is a commonly used approach that divides the recording into overlapping segments and applies a window function to each segment to reduce spectral leakage. A Fourier transform is then used to obtain each segment's frequency representation, from which a power estimate, called a modified periodogram, is calculated. These estimates are averaged to obtain the PSD, reducing variability compared with a single periodogram (Welch, 1967). When EEG amplitude is measured in microvolts, PSD is expressed in $\mu V^2/Hz$.

EEG power is commonly examined within frequency bands, including delta (approximately 1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (above 30 Hz). Because band boundaries vary across studies, the frequency ranges used should be reported explicitly. Band power can be expressed in absolute or relative terms. Absolute band power represents the total power within a specified frequency band and is calculated by integrating the PSD across that band. When EEG amplitude is measured in microvolts, absolute band power is expressed in μV^2. Relative band power expresses this value as a proportion of the total power across a specified reference frequency range that includes the band of interest. It is dimensionless and can also be reported as a percentage. The reference frequency range should therefore be stated explicitly.

Delta Activity

The delta band spans approximately 1–4 Hz. Studies examining this frequency range during meditation have reported regional reductions as well as no significant overall differences between recording conditions. (Cahn et al., 2010) reported reduced frontal delta power during Vipassana meditation relative to a mind-wandering control condition in an electrode-level analysis, particularly among participants who did not report drowsiness. However, the corresponding conventional ANOVA narrowly missed statistical significance, indicating that the evidence depended on the analysis used. (Young et al., 2021) found no significant difference in scalp-averaged, background-normalized delta power between meditation and mind wandering in experienced practitioners from six meditation traditions. This pooled result limits conclusions specific to mindfulness and does not exclude regional differences. (Duda et al., 2024) similarly reported no significant difference in traditional delta-band amplitude between eyes-closed mindfulness meditation and rest. These findings concern different spatial summaries and spectral measures and therefore should not be treated as equivalent evidence for a general increase or decrease in delta activity.

Theta Activity

The theta band is commonly defined as the frequency range from 4 to 8 Hz. In mindfulness research, theta power has been examined during meditation, immediately after brief exercises, and during cognitive tasks following training with studies reporting increases, decreases, and no consistent changes. These recording contexts address different questions about activity accompanying practice and changes observed outside meditation sessions.

During practice, (Nien et al., 2023) observed higher frontal theta power during both mindfulness and relaxation inductions than

during a control condition, indicating that this response was not specific to mindfulness. By contrast, (Hauswald et al., 2015) reported reduced frontal theta power in frontal during Zen open-monitoring meditation relative to rest.

Recordings surrounding brief exercises provide another perspective. (Susam et al., 2022) examined resting-state EEG before and after a two-minute mindfulness meditation exercise in 35 autistic youth. Total theta power averaged across electrodes was significantly higher afterward, whereas relative theta power showed no significant overall difference. This distinction illustrates that greater theta power does not necessarily mean a greater proportion of the reference spectrum. These findings concern immediate pre–post differences rather than changes across an extended training program.

(Nyhus et al., 2019) investigated theta during episodic memory retrieval before and after four weeks of mindfulness training, with a waitlist comparison group. The mindfulness group showed increased theta power over right frontal and left parietal sites following training, alongside improved source memory—the ability to remember the context in which information was encountered. However, (Luo et al., 2024) found no consistent pattern of midline theta change following a five-day mindfulness course. The evidence thus encompasses regional changes during practice, immediate differences after an exercise, and task-related changes following training, rather than a single general increase in theta activity.

Alpha Activity

Alpha findings, approximately 8–13 Hz, illustrate the importance of regional effects and recording procedures. (Ahani et al., 2014) reported higher alpha power primarily in right temporal and occipital regions during mindfulness meditation compared with podcast listening. (Luo et al., 2024) observed increased frontal

absolute alpha power during self-directed mindfulness practice following additional audio and video guidance. (Cahn et al., 2010) however, found no significant overall alpha-power difference between meditation and mind wandering in experienced Vipassana practitioners; alpha power was sensitive to the order of the recording conditions. Alpha power is also distinct from alpha peak frequency and alpha asymmetry. Whereas power quantifies activity within the selected band, peak frequency identifies the location of an alpha spectral peak, and asymmetry quantifies differences between corresponding left and right recording sites. These measures address different properties of alpha activity and should be reported separately.

Beta Activity

Beta findings, approximately 13–30 Hz, illustrate the distinction between activity during practice and changes across training. (Ahani et al., 2014) reported higher beta power during mindfulness meditation than during podcast listening in novice practitioners assessed following a six-week intervention. (Saggar et al., 2012) recorded EEG during mindfulness of breathing meditation at the beginning, middle, and end of a three-month intensive focused-attention retreat. Beta power decreased across training over anterior-central and posterior scalp regions on both sides of the head. A similar pattern emerged when the waitlist group subsequently completed the same training during a second retreat. These findings describe changes in activity recorded during meditation as training progressed.

Gamma Activity

Gamma activity, generally examined above 30 Hz, has attracted interest in studies of experienced meditators. (Cahn et al., 2010) reported increased parieto-occipital gamma power at 35–45 Hz during Vipassana meditation compared with mind wandering,

with stronger meditation-related increases among practitioners reporting more than ten years of daily practice. Similarly, increased posterior gamma power was observed during mindfulness compared to rest in (Berkovich-Ohana et al., 2012). Note that gamma recordings require particular attention to muscle artifacts, as electrical activity from scalp, facial, and neck muscles can contaminate high-frequency EEG signals as mention in (Whitham et al., 2007) .

Challenges in EEG Recording and Interpretation in Mindfulness Research

EEG findings in mindfulness research depend on how the practice, experimental comparison, and neural measurements are defined. Differences between studies may reflect these methodological choices as well as differences in the processes being investigated.

Practice Characteristics and Study Design

Mindfulness studies differ in their practice instructions, training duration, teaching context, and participants' previous experience. The label "mindfulness meditation" alone therefore does not establish that comparable practices were performed. Studies should describe the exercises, instructor qualifications, and methods used to assess practice and adherence (Davidson & Kaszniak, 2015).

The comparison condition also determines what can be concluded. Comparisons between experienced meditators and nonmeditators may reflect pre-existing differences or self-selection, whereas randomized longitudinal studies provide stronger evidence about training-related changes. Active controls help address shared influences, including expectations, instructor characteristics, group participation, and daily practice requirements (Davidson & Kaszniak, 2015) .

Recording Conditions and Signal Quality

Eye condition and alertness should be documented and kept comparable when they are not the experimental variables of interest. Researchers should also assess whether participants remained awake and followed the practice instructions. (Andrillon et al., 2021) identified sleep-like slow waves during wakefulness in association with attentional lapses, highlighting why slow-frequency changes during meditation require consideration of vigilance.

EEG recordings also contain non-neural activity, including signals from eye movements and muscles. Muscle contamination is particularly relevant to higher beta and gamma frequencies. et al. (Whitham et al., 2007) demonstrated substantial muscle contributions to scalp recordings above 20 Hz, indicating that high-frequency changes require careful evaluation before being attributed to neural activity. Studies should report artifact detection and correction procedures, the amount of data retained, and any differences in data quality between conditions or groups.

Measurement and Analytical Choices

Spectral findings depend on frequency boundaries, electrode selection, reference choice, and whether power is expressed in absolute or relative terms. Interpretation also requires distinguishing oscillatory peaks from the aperiodic spectral background. (Donoghue et al., 2020) showed that conventional band-power estimates can conflate these properties. An increase within a predefined frequency band therefore does not necessarily indicate that a distinct oscillation became stronger. Separating periodic and aperiodic contributions can help clarify the source of a reported difference.

Future Directions and Conclusion

Future EEG research can strengthen understanding of mindfulness by examining how specific practices relate to neural activity, how these relationships develop with training, and whether they accompany changes in everyday functioning. Clearly defined practice instructions and outcomes are essential for connecting EEG findings to these questions.

Repeated assessments before, during, and after an intervention could help establish when neural changes emerge and whether they persist. Including recordings during meditation and separate resting or cognitive tasks would allow researchers to examine whether changes extend beyond the practice session. Documenting practice frequency, duration, and participants' experiences would also help explain why individuals show different patterns of change.

A further priority is to test whether EEG changes relate to meaningful outcomes, such as improved attention, emotional regulation, or daily functioning. Studies should specify these proposed relationships in advance where possible and assess them directly. When neural and behavioral changes occur together, their association may help explain the significance of the EEG finding, although it does not by itself establish a causal mechanism.

EEG provides detailed information about the timing and frequency characteristics of brain activity during mindfulness practice and related tasks. The findings reviewed in this chapter show that its contribution depends on the question asked and the conditions under which activity is recorded.

References

Ahani, A., Wahbeh, H., Nezamfar, H., Miller, M., Erdogmus, D., & Oken, B. (2014). Quantitative change of EEG and respiration signals during mindfulness meditation. *Journal of NeuroEngineering and Rehabilitation*, *11*. https://doi.org/10.1186/1743-0003-11-87

Andrillon, T., Burns, A., Mackay, T., Windt, J., & Tsuchiya, N. (2021). Predicting lapses of attention with sleep-like slow waves. *Nature Communications* , *12*. https://doi.org/10.1038/s41467-021-23890-7

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*, 27–45. https://doi.org/10.1177/1073191105283504

Berkovich-Ohana, A., Glicksohn, J., & Goldstein, A. (2012). Mindfulness-induced changes in gamma band activity - Implications for the default mode network, self-reference and attention. *Clinical Neurophysiology*, *123*, 700–710. https://doi.org/10.1016/j.clinph.2011.07.048

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., Segal, Z. V., Abbey, S., Speca, M., Velting, D., & Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, *11*, 230–241. https://doi.org/10.1093/clipsy.bph077

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y. Y., Weber, J., & Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proceedings of the National*

Academy of Sciences of the United States of America, *108*, 20254–20259. https://doi.org/10.1073/pnas.1112029108

Brown, K. W., & Ryan, R. M. (2003). The Benefits of Being Present: Mindfulness and Its Role in Psychological Well-Being. In *Journal of Personality and Social Psychology* (Vol. 84, pp. 822–848). American Psychological Association Inc. https://doi.org/10.1037/0022-3514.84.4.822

Cahn, B. R., Delorme, A., & Polich, J. (2010). Occipital gamma activation during Vipassana meditation. *Cognitive Processing*, *11*, 39–56. https://doi.org/10.1007/s10339-009-0352-1

Crane, R. S., Brewer, J., Feldman, C., Kabat-Zinn, J., Santorelli, S., Williams, J. M. G., & Kuyken, W. (2017). What defines mindfulness-based programs? the warp and the weft. In *Psychological Medicine* (Vol. 47, pp. 990–999). Cambridge University Press. https://doi.org/10.1017/S0033291716003317

Davidson, R. J., & Kaszniak, A. W. (2015). Conceptual and methodological issues in research on mindfulness and meditation. *American Psychologist*, *70*, 581–592. https://doi.org/10.1037/a0039512

Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., Noto, T., Lara, A. H., Wallis, J. D., Knight, R. T., Shestyuk, A., & Voytek, B. (2020). Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, *23*, 1655–1665. https://doi.org/10.1038/s41593-020-00744-x

Duda, A. T., Clarke, A. R., & Barry, R. J. (2024). Mindfulness meditation alters neural oscillations independently of

arousal. *International Journal of Psychophysiology*, *205*. https://doi.org/10.1016/j.ijpsycho.2024.112439

Farb, N. A. S., Segal, Z. V., Mayberg, H., Bean, J., Mckeon, D., Fatima, Z., & Anderson, A. K. (2007). Attending to the present: Mindfulness meditation reveals distinct neural modes of self-reference. *Social Cognitive and Affective Neuroscience*, *2*, 313–322. https://doi.org/10.1093/scan/nsm030

Hauswald, A., Übelacker, T., Leske, S., & Weisz, N. (2015). What it means to be Zen: Marked modulations of local and interareal synchronization during open monitoring meditation. *NeuroImage*, *108*, 265–273. https://doi.org/10.1016/j.neuroimage.2014.12.065

Hölzel, B. K., Carmody, J., Vangel, M., Congleton, C., Yerramsetti, S. M., Gard, T., & Lazar, S. W. (2011). Mindfulness practice leads to increases in regional brain gray matter density. *Psychiatry Research - Neuroimaging*, *191*, 36–43. https://doi.org/10.1016/j.pscychresns.2010.08.006

Kabat-Zinn, J. (2003). Mindfulness-based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice*, *10*, 144–156. https://doi.org/10.1093/clipsy.bpg016

Kral, T. R. A., Davis, K., Korponay, C., Hirshberg, M. J., Hoel, R., Tello, L. Y., Goldman, R. I., Rosenkranz, M. A., Lutz, A., & Davidson, R. J. (2022). Absence of structural brain changes from mindfulness-based stress reduction: Two combined randomized controlled trials. *Science Advances*, *8*. https://doi.org/10.1126/sciadv.abk3316

Lazar, S. W., Kerr, C. E., Wasserman, R. H., Gray, J. R., Greve, D. N., Treadway, M. T., McGarvey, M., Quinn, B. T., Dusek, J. A., Benson, H., Rauch, S. L., Moore, C. I., & Fischl, B. (2005). Meditation experience is associated with increased cortical thickness. *NeuroReport*, *16*, 1893–1897. https://doi.org/10.1097/01.wnr.0000186598.66243.19

Lee, D. J., Kulubya, E., Goldin, P., Goodarzi, A., & Girgis, F. (2018). Review of the neural oscillations underlying meditation. In *Frontiers in Neuroscience* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fnins.2018.00178

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157. https://doi.org/10.1038/35084005

Lomas, T., Ivtzan, I., & Fu, C. H. Y. (2015). A systematic review of the neurophysiology of mindfulness on EEG oscillations. In *Neuroscience and Biobehavioral Reviews* (Vol. 57, pp. 401–410). Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2015.09.018

Loucks, E. B., Crane, R. S., Sanghvi, M. A., Montero-Marin, J., Proulx, J., Brewer, J. A., & Kuyken, W. (2022). Mindfulness-Based Programs: Why, When, and How to Adapt? *Global Advances In Health and Medicine*, *11*. https://doi.org/10.1177/21649561211068805

Luo, M., Gan, Q., Huang, Z., Jiang, Y., Li, K., Wu, M., Yang, D., Shao, H., Chen, Y., Fu, Y., & Chen, Z. (2024). Changes in Mental Health and EEG Biomarkers of Undergraduates Under Different Patterns of Mindfulness. *Brain*

Topography, *37*, 75–87. https://doi.org/10.1007/s10548-023-01026-y

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. In *Trends in Cognitive Sciences* (Vol. 12, pp. 163–169). https://doi.org/10.1016/j.tics.2008.01.005

Moore, A., Gruber, T., Derose, J., & Malinowski, P. (2012). Regular, brief mindfulness meditation practice improves electrophysiological markers of attentional control. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2012.00018

Nien, J. T., Gill, D. L., Chou, T. Y., Liu, C. S., Geng, X., Hung, T. M., & Chang, Y. K. (2023). Effect of brief mindfulness and relaxation inductions on anxiety, affect and brain activation in athletes. *Psychology of Sport and Exercise*, *67*. https://doi.org/10.1016/j.psychsport.2023.102422

Nyhus, E., Engel, W. A., Pitfield, T. D., & Vakkur, I. M. W. (2019). Increases in Theta Oscillatory Activity During Episodic Memory Retrieval Following Mindfulness Meditation Training. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00311

Saggar, M., King, B. G., Zanesco, A. P., MacLean, K. A., Aichele, S. R., Jacobs, T. L., Bridwell, D. A., Shaver, P. R., Rosenberg, E. L., Sahdra, B. K., Ferrer, E., Tang, A. C., Mangun, G. R., Alan Wallace, B., Miikkulainen, R., & Saron, C. D. (2012). Intensive training induces longitudinal changes in meditation state-related EEG oscillatory activity. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2012.00256

Slagter, H. A., Lutz, A., Greischar, L. L., Francis, A. D., Nieuwenhuis, S., Davis, J. M., & Davidson, R. J. (2007). Mental training affects distribution of limited brain resources. *PLoS Biology*, *5*, 1228–1235. https://doi.org/10.1371/journal.pbio.0050138

Susam, B. T., Riek, N. T., Beck, K., Eldeeb, S., Hudac, C. M., Gable, P. A., Conner, C., Akcakaya, M., White, S., & Mazefsky, C. (2022). Quantitative EEG Changes in Youth With ASD Following Brief Mindfulness Meditation Exercise. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 2395–2405. https://doi.org/10.1109/TNSRE.2022.3199151

Tang, Y. Y., & Posner, M. I. (2013). Tools of the trade: Theory and method in mindfulness neuroscience. *Social Cognitive and Affective Neuroscience*, *8*, 118–120. https://doi.org/10.1093/scan/nss112

Welch, P. D. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*, 70–73. https://doi.org/10.1109/TAU.1967.1161901

Whitham, E. M., Pope, K. J., Fitzgibbon, S. P., Lewis, T., Clark, C. R., Loveless, S., Broberg, M., Wallace, A., DeLosAngeles, D., Lillie, P., Hardy, A., Fronsko, R., Pulbrook, A., & Willoughby, J. O. (2007). Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clinical Neurophysiology*, *118*, 1877–1888. https://doi.org/10.1016/j.clinph.2007.04.027

Young, J. H., Arterberry, M. E., & Martin, J. P. (2021). Contrasting Electroencephalography-Derived Entropy and Neural Oscillations With Highly Skilled Meditators. *Frontiers in Human Neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.628417

CHAPTER 4

THE EVOLUTION OF HEALTHCARE NATURAL LANGUAGE PROCESSING FROM 2010 TO 2026

ERKAN YAŞAR[1]

Introduction

Clinical records contain histories, progress notes, discharge summaries, imaging reports, pathology narratives, patient messages, consent documents, and guideline text. These materials preserve context that structured fields often omit, but their heterogeneous language, abbreviations, temporal references, and institutional conventions make them difficult to analyze reliably. Healthcare natural language processing (NLP) emerged to convert such text into usable representations for research, quality measurement, coding, surveillance, and care delivery. Over time, the field has expanded from task-specific extraction pipelines to foundation-model systems that generate, retrieve, perceive, and act across clinical workflows.

This transition has changed both the technical object of healthcare NLP and the boundaries of the systems in which it

[1] Asst. Prof. Dr., Hakkari University, Faculty of Engineering, Department of Computer Engineering, Hakkari, Türkiye: 0000-0001-7333-6320

operates. Early work generally defined a narrow information problem, such as identifying a medication, detecting a symptom, assigning a phenotype, or removing identifiers. Developers constructed rules, lexicons, sectionizers, and statistical classifiers around that problem. Contemporary systems can summarize a longitudinal record, answer a clinical question using retrieved evidence, draft documentation from a conversation, or coordinate tools in a multi-step workflow. The technical opportunity is broader, yet the burden of validation is correspondingly larger: evaluating an extracted entity or a classification label is not equivalent to assessing a system whose outputs may influence documentation, communication, or clinical decisions.

This chapter uses the term *healthcare NLP* to include NLP applied to clinical documentation, biomedical literature, patient-facing communication, administrative text, and linked multimodal clinical information. It distinguishes this scope from *biomedical NLP*, which often emphasizes scientific literature and molecular knowledge. The distinction matters because a model that performs well on biomedical abstracts may still fail on shorthand, copied text, uncertainty, negation, and temporal ambiguity in clinical notes. Consequently, advances in general or biomedical language modeling cannot be assumed to translate directly into reliable performance in clinical settings.

The chapter traces the development of healthcare NLP through five overlapping phases: rule-based and feature-engineered clinical NLP (2010–2015); shared tasks, distributed representations, and reproducible data infrastructure (2013–2018); domain-adapted transformers (2018–2021); generative medical language models and retrieval-augmented generation (2022–2024); and multimodal, agentic, and prospectively evaluated systems (2024–2026). These periods are deliberately overlapping rather than mutually exclusive. Methods, datasets, and deployment practices advanced at different

rates, and newer approaches did not simply replace earlier ones. This chronology provides a framework for comparing landmark model families, open and controlled-access data resources, benchmarks, and evaluation frameworks while examining how their interactions shaped the field.

Three claims guide the discussion. First, progress has arisen through the interaction of representations, data infrastructure, and evaluation practices, rather than through model architecture alone. Second, clinical benchmark performance and clinical utility are different constructs. Benchmark accuracy is useful for measuring capabilities, but it does not establish whether a system improves care or supports a particular workflow safely and effectively. Clinical utility also depends on workflow fit, information provenance, calibration, equity, privacy, human oversight, and prospective outcomes. Third, the most consequential systems from 2024 onward are compound systems: they combine a model with retrieval, tools, interfaces, institutional data controls, and human decision makers. Their performance and safety must therefore be assessed at the system level, not inferred solely from the capabilities of the underlying model.

Retrieval-augmented generation (RAG) illustrates this broader shift. As a system design pattern, RAG can connect a language model to current evidence and local policy, making the retrieval process and the relationship between source material and generated output central to evaluation. However, retrieval does not by itself establish reliability. It introduces its own retrieval, attribution, and governance failure modes, which must be considered alongside errors in generation. Similar considerations apply to multimodal and agentic systems, whose expanded capabilities increase the importance of evaluating how models, information sources, tools, and human oversight interact within clinical workflows.

Taken together, these developments motivate the central argument of this chapter: the next stage of healthcare NLP should be assessed as a sociotechnical intervention rather than as a standalone language model. Understanding its trajectory therefore requires attention not only to what models can do, but also to the data, evidence, institutional arrangements, and evaluation practices that determine whether those capabilities become useful and safe components of healthcare.

Scope and Method of Historical Synthesis

This is a critical historical synthesis, not a systematic review or a meta-analysis. The chapter organizes the literature by methodological transition and treats dates as approximate periods of influence. A paper can belong to more than one phase, because datasets remain useful after a new architecture appears and older methods remain appropriate in constrained workflows.

The evidence base prioritizes publications and preprints with identifiable bibliographic records that illuminate one of four dimensions: model design, data resources, benchmark design, or clinical implementation. Preprints are explicitly identified as such. The discussion avoids pooling numerical results across datasets because task definitions, patient populations, annotation protocols, prevalence, and metrics differ. When a result is described as an improvement, that description concerns the study's own evaluation setting rather than a universal ranking.

The chapter also separates three layers of evidence. **Capability evidence** tests whether a system can complete a defined task on a benchmark. **Operational evidence** tests whether it integrates safely and efficiently into a workflow. **Clinical utility evidence** tests whether its use changes meaningful outcomes, such as documentation burden, timeliness, decision quality, equity, or

patient outcomes. These layers should be interpreted sequentially rather than substituted for one another.

Phase 1, 2010–2015: Rule-Based and Feature-Engineered Clinical NLP

At the beginning of the period, clinical NLP was dominated by task-specific pipelines. A typical system segmented documents, recognized terms from lexicons or dictionaries, detected negation and experiencer, normalized mentions to a terminology, and applied rules or supervised classifiers to produce a target label. These pipelines were closely tied to a corpus, note type, and institutional documentation style. Their components were often interpretable, which made it possible to inspect why a label had been produced and to modify behavior for a known local error.

Shared-task culture and annotated corpora were central to this phase. The field learned to define a task precisely, develop annotation guidelines, publish an evaluation metric, and compare systems on a held-out set. De-identification, concept extraction, relation extraction, temporal reasoning, and cohort identification became recurring problem formulations. The resulting benchmarks were valuable because they turned clinical language problems into reproducible engineering targets. They also narrowed the target: a strong score on a de-identified challenge dataset did not guarantee portability to a different health system.

Rule-based methods were not merely a prelude to machine learning. They were useful where clinical policy can be expressed explicitly, the vocabulary is relatively stable, and failure explanations are required. A carefully audited rule for a high-precision safety alert may be preferable to a general model with uncertain behavior. Yet manual rules scale poorly across note types and institutions. They also struggle with latent context, paraphrase, and interactions among distant spans of text.

The data environment was a second constraint. Clinical notes are sensitive personal data; their use requires governance, access controls, and privacy protection. The release of MIMIC-III created a major resource for reproducible critical-care research by making a deidentified database available under controlled Access (Johnson et al., 2016). Its importance was infrastructural as much as algorithmic: it enabled common tasks and promoted a research ecosystem around clinical data. It did not remove the external-validity problem, because a single critical-care database cannot represent all patients, documentation practices, or institutions.

Biomedical question answering also began to formalize evaluation during this period. BioASQ organized large-scale semantic indexing and question answering around biomedical information needs (Balikas, Krithara, Partalas, & Paliouras, 2015). Such challenges broadened the field beyond extraction, but they primarily measured answer-oriented capabilities under defined corpus conditions. Their relevance to bedside use depends on whether a system can locate current, applicable, and attributable information in a real clinical context.

Phase 2, 2013–2018: Representation Learning, Shared Resources, and Scalable Prediction

The second phase overlapped with the first. Sparse bag-of-words features, manually engineered lexical features, and rules increasingly coexisted with distributed word representations and neural sequence models. Representation learning reduced the need to enumerate every lexical variant, and recurrent or convolutional models supported end-to-end learning for selected tasks. Nevertheless, clinical applications still depended heavily on preprocessing, label construction, and dataset curation.

This period established a practical lesson: model selection follows task formulation. For a narrowly defined phenotype, a

transparent rule set and a simple classifier may be sufficient. For nuanced mention detection in varied narrative text, learned representations can improve recall and reduce manual feature design. A comparison of two NLP approaches for identifying bleeding among critically ill patients illustrates the continuing value of evaluating methods against a concrete clinical target rather than against architecture fashion alone (Taggart et al., 2018).

Data resources expanded the scale of retrospective work, while access conditions remained an essential part of research design. MIMIC-III coupled rich clinical records with structured variables and became a common substrate for mortality prediction, phenotyping, note classification, and representation learning (Johnson et al., 2016). Open biomedical resources, by contrast, often centered on scientific literature. The distinction created a persistent transfer gap. A model trained on formal research prose learns scientific terminology and discourse, whereas clinical notes contain compressed grammar, local templates, copy-forward artifacts, and patient-specific chronology.

By the end of the phase, the field had better corpora and more powerful representations, but evaluation usually remained task centered. This was appropriate for foundational measurement. It became insufficient once systems began to present generated or synthesized information to clinicians, because such outputs could influence decisions even when no single extraction label was visibly wrong.

Phase 3, 2018–2021: Domain-Adapted Transformers

Transformer pretraining changed healthcare NLP by turning language representation into a transferable resource. Instead of training a separate model from scratch for each dataset, researchers pretrained a model on large corpora and fine-tuned it for downstream tasks. The important methodological question became which corpus

and objective produce representations that transfer to a target domain.

BioBERT was an early landmark in this transition. It adapted a transformer language model to biomedical text and evaluated the resulting representations on biomedical text-mining tasks (Lee et al., 2019). The cited version is a preprint, so its status should be read accordingly. BioBERT demonstrated the value of domain-adaptive pretraining, but it also highlighted the difference between biomedical and clinical language. Scientific abstracts and clinical notes share terminology while differing in register, structure, and purpose.

ClinicalBERT and publicly available clinical BERT embeddings pushed adaptation closer to clinical records. ClinicalBERT focused on modeling clinical notes and predicting hospital readmission (Huang, Altosaar, & Ranganath, 2019), while the public clinical embeddings work made clinical-domain transformer representations available for downstream investigation (Alsentzer et al., 2019). Both cited reports are preprints. Their broader contribution was to establish clinical-domain pretraining as a standard design choice, while exposing practical limits: clinical text is hard to release, local corpora vary, and a pretraining source can encode institutional habits that do not transfer.

PubMedQA provided a benchmark for biomedical research question answering, with questions grounded in PubMed abstracts (Jin, Dhingra, Liu, Cohen, & Lu, 2019). It illustrates a recurrent pattern in healthcare NLP: a benchmark may be highly useful for comparing models yet still be distant from clinical decision support. Questions are curated, the evidence corpus is bounded, and the target answer format is defined. A clinician using a decision-support system faces incomplete records, ambiguity about the patient, changing guidelines, and responsibility for action.

Domain-specific pretraining was systematized by the study introducing biomedical language models trained from domain-relevant corpora, including models commonly known as PubMedBERT (Gu et al., 2020). The cited report is a preprint. Its contribution was methodological rather than a claim that one model family solves all healthcare tasks: corpus composition and tokenizer vocabulary influence what a model can represent. For clinical NLP, the implication is that pretraining choices should be reported as part of the model specification, alongside downstream labels and evaluation cohorts.

Table 1. Overlapping phases in healthcare NLP, 2010–2026

Phase	Approximate period	Dominant unit of work	Representative contribution	Principal limitation
1. Rules and features	2010–2015	Named entities, relations, phenotypes, de-identification	Explicit rules, lexicons, task-specific classifiers	Labor-intensive adaptation and limited semantic coverage
2. Representations and resources	2013–2018	Sequence labeling and prediction over expanding corpora	Neural representations and reproducible data resources	Task-centric evaluation and institutional transfer gaps
3. Domain transformers	2018–2021	Fine-tuned pretrained encoders	Biomedical and clinical transformer pretraining	Static knowledge and data-access constraints
4. Generative and retrieval systems	2022–2024	Question answering, summarization, drafting, evidence synthesis	Generative models and RAG pipelines	Hallucination, provenance, and evaluation mismatch

5. Multimodal and agentic workflows	2024–2026	Compound workflows across text, images, tools, and people	Multimodal foundation models and prospective workflow studies	Governance, interoperability, and prospective safety evidence

Phase 4, 2022–2024: Generative Medical Language Models and Retrieval-Augmented Generation

The generative turn altered the interaction model of healthcare NLP. Encoder-based transformers usually map input text to labels, spans, or embeddings. Generative models produce text and can therefore support question answering, summarization, drafting, explanation, and dialogue. This flexibility is attractive in healthcare because many clinical tasks end in narrative communication. It also creates a different failure surface: a fluent output may appear plausible even when it lacks support in the source record.

BioGPT exemplified the application of generative pretraining to biomedical text generation and mining (Luo et al., 2022). Its work made clear that generation should be evaluated beyond surface fluency. In a biomedical setting, an output may be grammatical and domain appropriate while still being factually unsupported, clinically irrelevant, or too uncertain for its wording. Generation therefore requires evidence linkage and task-specific review criteria.

Med-PaLM, reported as a large language model that encodes clinical knowledge, marked a widely visible move toward medical examination-style assessment and expert evaluation of generated answers (Singhal et al., 2023). The paper demonstrated that a general-purpose large language model adapted or prompted for medicine could perform strongly on selected medical question-answering evaluations. This is capability evidence. It should not be

interpreted as evidence that an autonomous system can safely diagnose, triage, or prescribe in routine care. Medical examinations contain valuable knowledge tests, yet clinical care also requires patient-specific evidence, communication, uncertainty management, and accountability.

RAG emerged as a response to the knowledge and provenance limitations of parametric generation. A RAG system typically indexes a curated collection, retrieves documents or passages for a query, and supplies them to a language model during answer generation. In clinical settings, the collection may include guidelines, institutional pathways, formularies, local policies, or the patient record. Almanac framed retrieval-augmented language models for clinical medicine and made the case that retrieved information can ground medical responses (Zakka et al., 2023). The cited report is a preprint.

RAG should be understood as a pipeline rather than a property of a language model. Its quality depends on source curation, indexing, query formulation, retrieval recall, passage selection, context-window allocation, generation constraints, citation display, and user interface. Retrieval can reduce unsupported claims when the relevant evidence is found and faithfully used. It can also fail silently when the corpus is incomplete, the retriever selects irrelevant evidence, a source is out of date, or the generator overstates what a passage supports.

Table 2. Architectural transition and suitable evaluation focus

Approach	Typical output	Strength in healthcare workflows	Core evaluation focus	Common failure mode
Rules and terminology pipelines	Concept or alert	Traceable policy implementation	Precision, recall, error audit, local maintainability	Coverage loss under linguistic variation

Task-specific neural model	Label, span, or score	Learned patterns at scale	External validation, calibration, subgroup performance	Dataset shift and opaque shortcuts
Domain-adapted encoder	Transferable representation	Efficient fine-tuning across related tasks	Task fit, corpus provenance, portability	Mismatch between pretraining and clinical text
Generative language model	Narrative answer or draft	Interactive synthesis and documentation support	Factuality, completeness, uncertainty, human review	Plausible unsupported generation
RAG system	Answer linked to retrieved sources	Current and local knowledge integration	Retrieval recall, attribution fidelity, source currency	Relevant evidence absent or misused
Agentic workflow	Multi-step action plan or tool-mediated output	Orchestration across systems	Step-level auditability, authorization, end-to-end safety	Compounded tool and handoff errors

Phase 5, 2024–2026: Multimodal, Ambient, and Agentic Workflows

From 2024, the unit of innovation increasingly became a workflow rather than a single model. Multimodal systems connect text with images and other clinical data modalities. Ambient systems turn clinician–patient conversation into documentation artifacts. Agentic systems decompose a request into multiple steps, call retrieval or software tools, and maintain intermediate state. These configurations can address real workflow fragmentation, but they require evaluation at the level of the entire sociotechnical system.

A multimodal biomedical foundation model trained from fifteen million image–text pairs illustrates the scale and breadth of

contemporary image–language pretraining (Zhang et al., 2024). Multimodal pretraining can create representations that relate visual findings to textual descriptions. Clinical use still depends on modality-specific validation, dataset provenance, label fidelity, and a clear statement of the intended role. An image–text representation model is not equivalent to a validated diagnostic device, and its behavior can change when imaging protocols, populations, or reporting styles shift.

Ambient AI scribes provide a useful example of workflow-centered evaluation. The study of ambient scribes aimed to alleviate clinical documentation burden (Tierney et al., 2024). In this setting, accuracy is necessary but not sufficient. An omitted symptom, a fabricated finding, or an incorrectly attributed statement can affect the clinical record. Evaluation therefore needs documentation completeness, attribution, edit burden, privacy safeguards, clinician acceptance, and monitoring for downstream consequences. The relevant counterfactual is the existing documentation process, not an idealized zero-error transcript.

Agentic language-model workflows remain an emerging design pattern. In healthcare, a safe interpretation is narrow: an agent can assist with decomposition, retrieval, drafting, or routing under explicit permissions and human review. The term should not imply independent clinical authority. Each tool invocation can introduce errors in authentication, patient selection, parameter interpretation, source retrieval, and state management. A system that combines multiple individually acceptable components can fail at their interfaces.

Prospective clinical evaluation began to appear more prominently in this period. A 2026 Nature Medicine report described prospective evaluation of a large-language-model clinical decision-support system in an emergency department (Leibovitch et al., 2026). Such studies are especially important because they evaluate a

system in the context where outputs are seen and acted upon. Their conclusions must nevertheless be read with attention to setting, comparator, supervision, case mix, and endpoints. Prospective evaluation is stronger evidence of operational relevance than retrospective benchmark performance, yet it does not automatically generalize across institutions.

Data Resources, Benchmark Design, and the Meaning of Generalization

Datasets determine what a model can learn and what an evaluation can claim. Healthcare datasets differ across access regime, documentation source, annotation target, clinical setting, language, and linkage to outcomes. These differences make a single global leaderboard misleading. A dataset may be carefully annotated yet narrow in patient population; another may be large but weakly labeled; a third may be clinically representative but inaccessible for reproducible comparison.

MIMIC-III remains an influential example of a controlled-access clinical database that enabled reproducible research while preserving a governance framework for sensitive data (Johnson et al., 2016). Its critical-care focus supports important research questions but should not be treated as a generic proxy for all clinical text. Systematic examination of clinical text datasets for medical AI and large language models further emphasizes that access, corpus composition, and annotation conditions shape the available evidence (Wu et al., 2024).

Benchmark design contains normative choices. A de-identification benchmark prioritizes privacy-sensitive identifiers. A concept-extraction benchmark prioritizes span and concept agreement. A medical question-answering benchmark prioritizes an answer keyed to a reference response. An ambient documentation

evaluation prioritizes whether a note supports real work. None of these objectives subsumes the others. A mature evaluation plan starts by identifying the decision or workflow that the system will affect, then chooses measures aligned with that use.

Generalization has at least four meanings in this domain. **Temporal generalization** asks whether performance persists as practice and documentation change. **Geographic generalization** asks whether a system transfers across institutions. **Population generalization** asks whether performance is equitable across relevant patient groups. **Task generalization** asks whether a model can be repurposed without hidden degradation. Foundation models often appear broad on the fourth definition, while clinical deployment requires evidence on all four.

Table 3. Selected data and benchmark resources discussed in this chapter

Resource	Primary content	Research value	Interpretive boundary
MIMIC-III	Deidentified critical-care clinical database	Reproducible retrospective modeling and clinical NLP research (Johson et al., 2016)	One data-generating environment; not a universal clinical-text distribution
BioASQ	Biomedical semantic indexing and question answering	Structured evaluation of biomedical information tasks (Balikas, Krithara, Partalas, & Paliouras , 2015)	Biomedical information-seeking differs from patient-specific care
PubMedQA	Biomedical research questions grounded in abstracts	Question-answering evaluation over research literatüre (Jin, Dhingra, Liu, Cohen, & Lu, 2019)	Bounded evidence and curated questions limit bedside inference

Clinical text dataset landscape	Heterogeneous clinical-text corpora and access arrangements	Frames data availability as a methodological and governance issue (Wu et al, 2024)	Dataset inventories do not establish cross-dataset comparability
Multimodal image–text corpus	Large-scale biomedical image–text pairs	Pretraining for biomedical multimodal representations (Zhang et al., 2024)	Pretraining scale does not substitute for clinical validation

Applications: From Extraction to Documentation and Decision Support

Healthcare NLP applications can be organized by the relationship between output and action. Extraction systems identify information from text for registries, quality measurement, research cohorts, and operational analytics. Documentation systems draft, summarize, or structure narrative content. Decision-support systems surface evidence, synthesize record information, or propose possibilities that a clinician reviews. Administrative systems support coding, authorization, inbox routing, and communication. Each category has distinct tolerances for error.

For extraction, the central questions are label definition, annotation reliability, and portability. A model that identifies a condition from a particular note section may work well for retrospective cohort selection yet fail as a real-time trigger if the relevant text is documented later. For documentation, the key risk is record integrity: generated language can become part of the legal and clinical record. For decision support, the key risk is undue reliance on a fluent system whose claims exceed its evidence.

Medical coding provides a concrete setting in which generative and retrieval-based systems may assist rather than replace accountable coders and clinicians. A 2025 study assessed retrieval-augmented large language models for medical coding (Klang et al., 2025). It is an example of a task where evaluation should include

coding correctness, evidence traceability, handling of uncertainty, workflow time, and review behavior. A coding score alone cannot establish whether a system improves revenue-cycle integrity or reduces compliance risk.

Clinical decision support faces a higher evidentiary threshold because outputs can alter care. The evaluation and mitigation study of large language models in clinical decision-making underscores the need to identify limitations before deployment (Hager et al., 2024). A safe design constrains the use case, identifies who reviews the output, presents source evidence, and defines what the system must decline to answer. It also treats updates to the model, retrieval corpus, prompt, and interface as changes requiring governance rather than as routine maintenance.

RAG as Clinical Information Infrastructure

RAG is frequently described as a way to reduce hallucination. That description is incomplete. RAG changes the information architecture of a system by shifting part of the answer from parametric memory to a governed corpus. This can be advantageous in medicine, where guidelines change, local pathways differ, and users often need to inspect the basis for an answer. It also introduces an operational dependency on retrieval quality and source governance.

A clinically oriented RAG system should define the corpus before choosing the generator. A guideline assistant requires versioned guideline documents, metadata for issuing organization and effective date, and clear treatment of conflicting recommendations. A patient-record assistant requires identity assurance, encounter boundaries, note provenance, and role-based access. An institutional policy assistant requires a process for retiring superseded policies. Without such controls, an apparently cited answer may still rely on stale, irrelevant, or inapplicable material.

At least six evaluation questions follow. First, did retrieval return evidence relevant to the query? Second, did it return the evidence needed to answer safely, rather than merely a topically related passage? Third, did the generation accurately represent the retrieved material? Fourth, are displayed citations attached to the claims they support? Fifth, can the user detect uncertainty or evidence conflict? Sixth, does the system behave safely when no adequate evidence is retrieved? Almanac provides an early clinical RAG framing (Zakka et al., 2023), and a later scoping review maps technical implementations, clinical applications, and ethical considerations in medical RAG (Yang et al., 2025). The latter is a preprint and should be interpreted as a field-mapping contribution rather than definitive deployment evidence.

Table 4. RAG evaluation layers for healthcare use

Layer	Question	Illustrative evidence	Why it matters
Corpus governance	Are sources authoritative, versioned, and in scope?	Source inventory, update logs, access policy	Retrieval cannot correct an incomplete or stale knowledge base
Retrieval	Was the needed evidence retrieved?	Recall-oriented review against adjudicated evidence	A correct generator cannot use material it never receives
Grounding	Does each claim follow from the retrieved source?	Claim–evidence annotation and attribution audit	Topical citation is weaker than faithful support
Generation	Is the response complete, calibrated, and comprehensible?	Expert review, uncertainty analysis, omission audit	Fluent language can conceal missing or overstated content
Workflow	Does use improve a real task without new burden?	Time, edits, overrides, usability, handoff observations	Technical quality may not translate into operational value

Safety and equity	Who is harmed when the system fails?	Subgroup analysis, incident monitoring, escalation pathway	Aggregate performance can hide concentrated risk

Evaluation, Safety, Equity, and Governance

Evaluation should be proportional to the plausible consequence of an output. A background literature assistant may be assessed initially through retrieval relevance and citation fidelity. A system that drafts a clinical note requires additional assessment of factual correspondence to the encounter, edit burden, and propagation of errors into the record. A system that informs triage or treatment decisions requires prospective, context-specific evidence with explicit human oversight and governance.

Benchmark metrics remain useful. Span-level precision and recall reveal extraction errors; classification metrics reveal discrimination; exact match or answer scoring reveals conformance to a benchmark target; human ratings can capture dimensions that automated metrics miss. None alone captures clinical utility. A high benchmark score can coexist with poor calibration, uneven subgroup performance, poor user interface design, or an unacceptable failure mode. Conversely, a modestly performing model can create value if it assists a bounded low-risk task and is easy to verify.

Safety evaluation for generative systems must include both commissions and omissions. An unsupported recommendation is a commission error. Failure to retrieve a relevant contraindication is an omission error. A response that states uncertainty clearly may be safer than one that gives an unqualified but incomplete answer. The mitigation study by Hager and colleagues emphasizes structured approaches to limitations in clinical decision-making settings (Hager et al., 2024).

Equity cannot be inferred from overall performance. Language models may encode patterns that disadvantage groups through differential data representation, stereotype reinforcement, language mismatch, or unequal access to system benefits. The health-equity toolbox for large language models offers a framework for surfacing such harms and biases (Pfohl et al., 2024). Equity evaluation should begin during problem formulation: researchers should ask which populations are represented in the development data, how labels were produced, who can use the resulting system, and what recourse exists when an output is harmful.

Governance must treat a deployed language-model system as a changing intervention. The model, prompt, retrieval index, source documents, access-control layer, and user interface can all change behavior. A governance program therefore needs versioning, audit logs, incident reporting, periodic re-evaluation, and clearly assigned accountability. These requirements apply equally to commercial and locally developed systems.

Table 5. Capability, operational performance, and clinical utility are distinct evaluation targets

Evaluation target	Core question	Typical measures	What it cannot establish alone
Technical capability	Can the system perform a defined task on a defined dataset?	Precision, recall, discrimination, answer correctness, factuality	Safe use in an institution or benefit to patients
External validity	Does performance persist across time, sites, and populations?	Temporal, geographic, and subgroup validation	Workflow adoption or causal impact
Operational performance	Can people use the system reliably in a workflow?	Time, edits, overrides, usability, failure recovery	Improvement in health outcomes

Clinical utility	Does use improve meaningful outcomes with acceptable harms?	Prospective endpoints, safety events, decision quality, burden	Generalizability beyond the studied implementation
Governance readiness	Can the system be monitored and controlled after release?	Auditability, versioning, access control, escalation processes	That the intended task is clinically beneficial

Persistent Gaps and Research Agenda

The central gap is the distance between benchmark construction and clinical implementation. Benchmarks require stable labels and bounded inputs. Clinical work involves changing records, incomplete evidence, interruptions, and teams with different responsibilities. Future research should connect these settings without abandoning the rigor that benchmarks provide.

First, the field needs stronger data documentation. Reports should specify data origin, access conditions, time period, note types, preprocessing, label construction, missingness, and known representational gaps. Second, external validation should become routine for systems proposed for clinical use. A model should be evaluated across institutions or at least across temporally separated cohorts before broad claims of robustness are made.

Third, evaluation must address compound systems. For a RAG assistant, it is insufficient to test only the language model. Researchers should test retrieval coverage, source currency, claim-level grounding, interface comprehension, and failure recovery. For an agentic workflow, they should test each tool boundary, authorization pathway, and state transition. The system-level result should be traceable to component-level evidence.

Fourth, prospective studies should use endpoints matched to intended use. Ambient documentation tools may be assessed through

documentation burden, note integrity, and clinician time. Coding tools may be assessed through coding quality, review workload, and compliance outcomes. Decision-support tools may require carefully staged studies of decision quality, safety, and patient-centered outcomes. The 2026 prospective emergency-department study points toward this higher standard of evidence (Leibovitch et al.,2026).

Fifth, healthcare NLP needs explicit boundaries for automation. The relevant question is rarely whether a model can produce an answer. It is whether the answer should be used, by whom, under what review, with what evidence, and with what recourse if it is wrong. Designing for refusal, escalation, and transparent uncertainty is as important as maximizing response completeness.

Conclusion

From 2010 to 2026, healthcare NLP evolved from manually engineered extraction pipelines to foundation-model ecosystems that can generate language, retrieve evidence, integrate modalities, and coordinate tools. The transition brought wider capability and new opportunities for documentation, information access, coding, and decision support. It also expanded the sources of error and the requirements for evidence.

The enduring lesson is that healthcare NLP should be judged by the work it changes. Rules, classifiers, transformers, generative models, RAG pipelines, and agents each have roles when their intended use, evidence, and safeguards are explicit. Benchmark results remain necessary for scientific progress, but clinical utility requires a further chain of evidence: external validity, workflow evaluation, safety analysis, equity assessment, governance, and prospective study. The most credible future systems will combine technical sophistication with disciplined limits on what their outputs are allowed to mean.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly available clinical BERT embeddings* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1904.03323

Balikas, G., Krithara, A., Partalas, I., & Paliouras, G. (2015). BioASQ: A challenge on large-scale biomedical semantic indexing and question answering. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-319-24471-6_3

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). *Domain-specific language model pretraining for biomedical natural language processing* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2007.15779

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*. https://doi.org/10.1038/s41591-024-03097-1

Huang, K., Altosaar, J., & Ranganath, R. (2019). *ClinicalBERT: Modeling clinical notes and predicting hospital readmission* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1904.05342

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A dataset for biomedical research question answering* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1909.06146

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., &

Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*. https://doi.org/10.1038/sdata.2016.35

Klang, E., Tessler, I., Apakama, D. U., Abbott, E., Glicksberg, B. S., Arnold, M., Moses, A., Sakhuja, A., Soroush, A., Charney, A. W., Reich, D. L., McGreevy, J., Gavin, N., Carr, B., Freeman, R., & Nadkarni, G. N. (2025). Assessing retrieval-augmented large language models for medical coding. *NEJM AI*. https://doi.org/10.1056/AIcs2401161

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). *BioBERT: A pre-trained biomedical language representation model for biomedical text mining* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1901.08746

Leibovitch, L., Ahituv, A., Gorenshtein, A., Aran, D., Sorka, M., Miron, K., & Shelly, S. (2026). Prospective evaluation of a large language model clinical decision support system in the emergency department. *Nature Medicine*. https://doi.org/10.1038/s41591-026-04601-5

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbac409

Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., Tomasev, N., Rashid, Q. M., Azizi, S., Rostamzadeh, N., McCoy, L. G., Celi, L. A., Liu, Y., Schaekermann, M., Walton, A., Parrish, A., Nagpal, C., Singh, P., Dewitt, A., . . . Singhal, K. (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*. https://doi.org/10.1038/s41591-024-03258-2

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*. https://doi.org/10.1038/s41586-023-06291-2

Taggart, M., Chapman, W. W., Steinberg, B. A., Ruckel, S., Pregenzer-Wenzler, A., Du, Y., Ferraro, J., Bucher, B. T., Lloyd-Jones, D. M., Rondina, M. T., & Shah, R. U. (2018). Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA Network Open*. https://doi.org/10.1001/jamanetworkopen.2018.3451

Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst*. https://doi.org/10.1056/CAT.23.0404

Wu, J., Liu, X., Li, M., Li, W., Su, Z., Lin, S., Garay, L., Zhang, Z., Zhang, Y., Zeng, Q., Shen, J., Yuan, C., & Yang, J. (2024). Clinical text datasets for medical artificial intelligence and large language models—A systematic review. *NEJM AI*. https://doi.org/10.1056/AIra2400012

Yang, R., Wong, M. Y. H., Li, H., Li, X., Zhu, W., Liao, J., Yu, K., Liew, J. C. K., Xuan, W., Chen, Y., Ke, Y., Ong, J. C. L., Teodoro, D., Hong, C., Ting, D. S. W., & Liu, N. (2025). *Retrieval-augmented generation in medicine: A scoping review of technical implementations, clinical applications, and ethical considerations* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.05901

Zakka, C., Chaurasia, A., Shad, R., Dalal, A. R., Kim, J. L., Moor, M., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Nelson, J., & Hiesinger, W. (2023). *Almanac: Retrieval-augmented language models for clinical medicine* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.01229

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., . . . Poon, H. (2024). A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*. https://doi.org/10.1056/AIoa2400640


BİDGE